ROBUST REMOTE SEISMIC STATION

Paul B. Akers and Kevin B. Wilbanks

Geotech Instruments, LLC

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. DE-FG02-04ER83966

ABSTRACT

The United States government sponsors and supports various programs to monitor nuclear explosions through seismic, hydroacoustic, and infrasound data collection stations. These stations are typically deployed in remote areas all over the world. However, current remote data acquisition station technology is limited by (1) poor data quality and reliability and (2) high installation, operation, and maintenance costs. This project was initiated to begin the development of a new generation of compact, remote, seismic data acquisition hardware and software using advanced low power electronics, packaging, and power source technologies. It specifically addresses the problems of data quality and data communications reliability and will reduce deployment, operational, and maintenance costs.

This paper describes the work conducted in Phase I of this project that investigated new technologies and evaluated the feasibility of using these technologies to successfully achieve this project's goals. Specifically Phase I investigated the following: (1) new low power high-resolution analog-to-digital converter (ADC) devices and configurations, (2) overall size and power reduction using low power highly integrated miniaturized components, (3) integrated most of the major subsystems of the station into a single electronics package, (4) new low power satellite and wireless commutations options, and (5) advanced power supplies such as fuel cell technologies. Phase I found that due to the convergence of these technologies driven by the consumer mass market, it was feasible to develop a much improved seismic data acquisition platform for future needs.

Continuing forward into Phase II, this project will result in a working prototype system for a new generation of robust remote seismic stations that will integrate the technologies investigated in Phase I. Using the conceptual designs developed in Phase I, prototype hardware and software will be designed, fabricated, and tested in Phase II. This project specifically pertains to land based borehole seismic data collection stations, however the concepts developed are equally applicable to surface vault type seismic, hydroacoustic, and infrasound data collection stations.

In addition to national security and nonproliferation applications, the hardware and software developed in this project will be directly applicable to commercial earthquake, strong motion, infrasound, and hydroacoustic data acquisition and monitoring. By providing high quality reliable data, more data will be available for monitoring and study activities that should bring about advances in these areas. Also, by providing lower overall installation, operation, and maintenance costs, more stations should be able to be fielded with higher productivity. All of this will improve seismic data acquisition coverage and monitoring capabilities that in turn will improve the assessment of nuclear nonproliferation and seismic hazards and reduce the associated risk of each.

OBJECTIVE

The United States government sponsors and supports various programs to monitor nuclear explosions by means of seismic, hydroacoustic, and infrasound data collection and analysis. These data collection stations are typically deployed in remote areas to remove them from culturally generated noise signals. Also, to perform their monitoring functions, they may need to be located in politically sensitive areas. These requirements give rise to several problems encountered with past and current station designs. These include the following:

- High installation, operation, and maintenance costs.
- High physical profile because of the equipment and infrastructure required to support the station.
- Low reliability of intrasite data communications in remote areas as well as communications back to a central data collection center.
- Power availability, consumption and generation.

The objective of this Small Business Innovation Research (SBIR) Phase I grant was to research and assess the feasibility of using current and emerging new technologies to develop a new generation of Robust Remote Seismic Station (RRSS) hardware and software to mitigate these current shortcomings. The working goals for this new generation of equipment are as follows:

1. High quality 24-bit or better data.
2. Remove the need for a central hub data collection point by having the individual stations connect directly to a global or wide-area network (GAN/WAN) to eliminate a large physical presence and cost. It also removes a large power requirement and a possible single point of failure.
3. Very low power (<5 W peak, <1 W average at each seismic station).
4. High level of integration to reduce size and provide a very low physical profile.
5. Direct low power low Earth orbit (LEO) satellite communication to “Internet in the sky.”
6. Two-week persistent data buffer in case of communications outages.
7. Autonomous power operation (off the grid).
8. One-year operation between servicing.

Specifically, Geotech Instruments, LLC (Geotech) performed research in the following areas:

- **High Quality Data**: Using new low power analog signal conditioning and high-resolution delta-sigma ($\Delta\sum$) ADC techniques to deliver high quality, low noise seismic data.

- **Size and Power Reduction**: Using low power surface mount electronics and miniaturized packaging technologies to reduce overall size and power requirements.

- **Integration**: The integration of the electronics modules into one borehole installable instrument package to allow the physical profile of the station to be reduced. By minimizing the total number of components and instrument packages, cabling and interconnects was minimized thus reducing installation, operation, and maintenance costs significantly. The ultimate goal was to achieve a near zero maintenance system that could be quickly deployed by inexperienced personnel. The system would be plug and play requiring few if any field configuration or adjustments.

- **Communications**: Using new low power and low cost satellite and wireless communications to connect each station directly back to a data center removing the need for a remote central data hub. This eliminated problems encountered with intrasite communications and a possible single point of failure. Eliminating the central hub also reduces installation, operation and maintenance costs. Data from the new seismic station will be delivered in standard CD 1.1 formats and protocols to maintain compatibility with and preserve the investment in current data center processing hardware and software.

- **Advanced Power Supplies**: Using new fuel cell technology as the power source for the station. By significantly reducing the power consumption of the station from current levels, new smaller power source options can be used. This allows for a large reduction in the installation, operation, and maintenance costs associated with the power subsystem.

RESEARCH ACCOMPLISHED

High Quality Data

The Phase I goals in this area were to investigate:

- Reduction of system induced analog noise.
- New lower power, low noise analog front-end components.
- New lower power, high-resolution ADC components.
- Discrete ΔΣ ADC and digital signal processor (DSP) processing methods.
- Matching sensors, analog front-end, and ADC for optimized performance.

Prior to proposing this SBIR program, Geotech's standard 24-bit data acquisition products were its line of D-Series instruments. These instruments used 1st generation ΔΣ ADC chips. The total power requirements for a three channel ADC (analog front-end, ΔΣ modulator and DSP finite impulse response [FIR] filter) was 1630 mW and required 43.7 sq in. of printed circuit (PC) board space.

Development of Geotech's newest line of 24-bit data acquisition products, the SMART-24™, and this SBIR program proposal phase and Phase I activities occurred concurrently. In this effort, 2nd generation ΔΣ ADC and analog front-end chips were designed in and evaluated. In this hardware, the total power requirements for a three channel ADC (analog front-end, ΔΣ modulator and DSP FIR filter) was drastically reduced to 258 mW and required only 13.5 sq in. of PC board space. It was also found that this generation of devices was significantly less sensitive to induced noise from the digital portions of the hardware (see Figure 1). This greatly reduced the amount of isolation, shielding and special PC board layout required to obtain good performance results. A 3–6 dB improvement in dynamic range was also achieved.

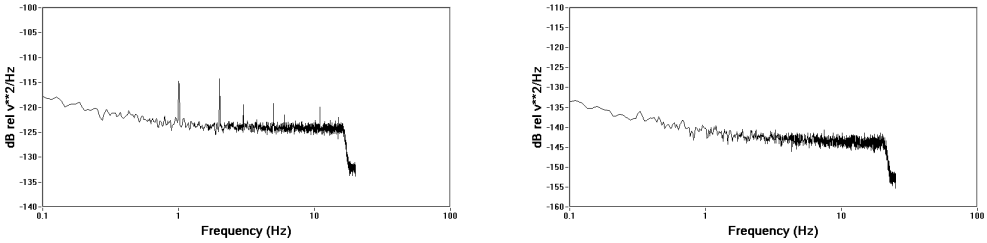


Figure 1. Input terminated noise of a 1st generation ADC (left, showing a small amount of digitally induced integer hertz noise) and a 2nd generation ADC (right, showing no digital pickup).

In evaluating this design further, it was found that additional improvements can be made by optimizing power supply voltages, matching the ADC to the sensor and by moving the DSP FIR filtering functions into underutilized field-programmable gate array (FPGA) and DSP processor resources that are consuming power but not being used effectively. This could reduce the power and PC board space even further to 150 mW and 4.7 sq in., respectively. These results are summarized in Table 1.

Table 1. Three Channel ΔΣADC Power and PC Board Space Requirements.

	1st Gen. ΔΣ ADC	2nd Gen. ΔΣ ADC	SBIR 3rd Gen. ΔΣ ADC
Analog Front End	940 mW	99 mW	75 mW
ΔΣ Modulator	630 mW	99 mW	75 mW
DSP Filter	60 mW	60 mW	N/A
Total Power	**1630 mW**	**258 mW**	**150 mW**
Board Space	43.7 $in.^2$	13.5 $in.^2$	4.7 $in.^2$

In addition, Geotech is currently cooperating with various vendors in the development of new 24-bit ΔΣ ADC devices that will be available in the late 2005 and 2006. These devices promise to further reduce power, board space and cost while improving the dynamic range performance and they will be evaluated more fully in Phase II of this project as they become available.

Discrete ΔΣ ADC topologies were also reviewed and studied in Phase I that have the potential to push ADC resolution past the 24-bit barrier. This evaluation will continue under Phase II where prototype circuits will be designed, built, and tested.

Size and Power Reduction

The Phase I goals in this area were to investigate the following:

- Reduction of PC board size using new miniaturized components (see Figure 2).
- Reducing power requirements to a minimum.

Given recent advancements in both miniaturized and low-power integrated circuits driven by the wireless phone, personal digital assistant (PDA), and digital camera industries, smaller, more-compact low-power electronics can be designed. While FPGAs typically use more quiescent power, the overall power requirement is lower if 1.8-V I/O, peripherals and System on a Programmable Chip (SoPC) modules can be used and integrated into a larger device. Many of the peripheral building blocks, DSPs, and microcontrollers used in Geotech's current SMART-24™ system can be integrated into a single FPGA to significantly reduce size and overall power. This reduced size will result in an overall reduction in cost, power, physical size, and weight. Currently, Geotech's borehole digitizers have a 3.5-in. diameter and are roughly 26 in. long. The reduction in size will result in a new generation of borehole digitizers, which will be targeted at a 3.5-in. diameter and a length of 8 in. This reduction in size will reduce the weight by an estimated 75%–80%.

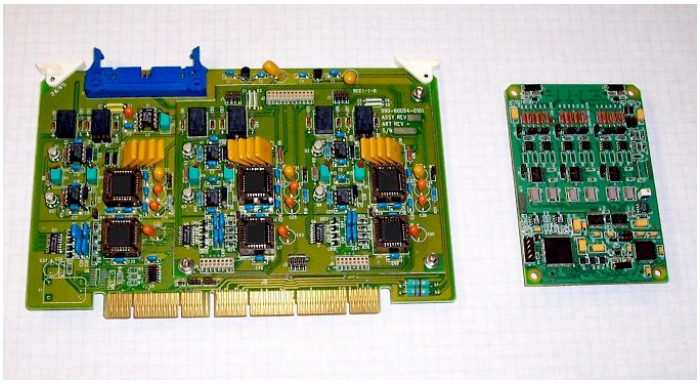


Figure 2. This figure shows a 70% reduction in PC board size from a 1st generation ADC (left) to a 2nd generation ADC (right). Another 50% reduction in size is anticipated going to a new 3rd generation of ADC.

Hardware developed in Phase II of this program will be the building block for Geotech's next generation surface digitizer to replace the SMART-24™ instruments carrying forward Geotech's intellectual property and innovations respected around the world. This investment in intellectual property will allow Geotech to advance its hardware platform as newer, faster, and lower-power integrated circuits come to market; all this with maximum portability via very high speed integrated circuit hardware description language design tools and methodologies. FPGAs where selected over application-specific integrated circuits (ASICs) because of a more rapid migration to newer technology and programmability. By letting the FPGA industry invest time and money into the semiconductor devices, Geotech can stay 2 to 3 years ahead of the ASIC based competition. The current borehole digitizer contains nine boards and consumes 188 sq in. of board space. Under this program, the next generation borehole digitizer is targeted for two boards and a 39-sq-in. area.

This integration of SoPCs onto a single FPGA will allow Geotech to lay the foundation for an integrated ΔΣ ADC, which could expand the range of the current 24-bit digitizers beyond the 24-bit limit. Having the core instantiated

internal to the FPGA, the ΔΣ ADCs and ΔΣ DACs can be dynamically scaled for lower speed ADCs and DACs. In seismology, sometimes 24-bit resolution is excessive. Lower resolution customers could benefit by the power savings of a 16-bit or 20-bit system; all this without investing in changes to the hardware design. Also, multiple 10-bit and 12-bit ΔΣ ADCs and ΔΣ DACs can be instantiated for state-of-health (SOH) type digital signaling.

Integration

The Phase I goals in this area were to investigate the following:

- Overall station size reduction to reduce its physical profile.
- The integration of components into a single package.
- Providing simple installation and user-friendly operation.
- Near zero maintenance and configuration, easy to identify failures and to repair.

Reduced PC board size requirements allow for more components of the station to be integrated into a single package. This will allow the pre-amplifiers, global positioning system (GPS) receiver, and data authentication to all be integrated into a single digitizer package with fewer boards and interconnects. This in turn will increase reliability and lower manufacturing and maintenance costs. Reducing the number of total components and interconnections will allow for a smaller overall physical profile of the station. Stations will be installed in such a way so that only the GPS and communications antennas are exposed to reduce its above ground profile (see Figure 3).

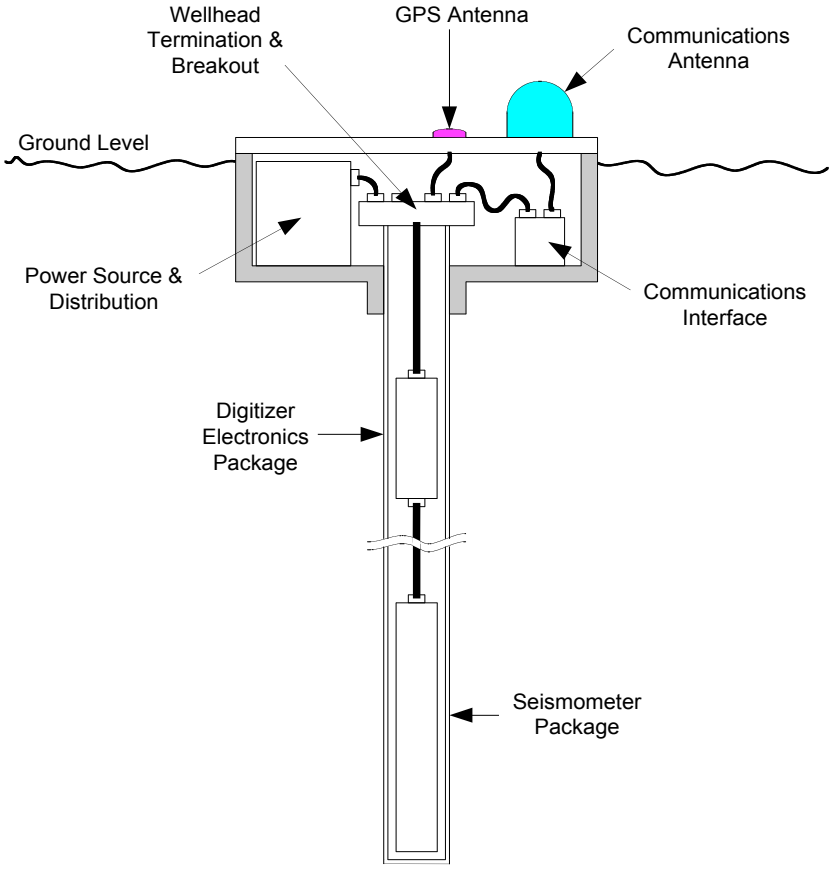


Figure 3. Typical RRSS system integration.

Because the individual stations can connect directly to a network, the need for a central data hub in an array is removed. This eliminates a large installation, maintenance, and power cost from the system. The use of SMART-24™ plug and play technology in the digitizer allows it to automatically detect the components connected

to it and to take the appropriate configuration actions. HTTP web, Telnet, and FTP access to each station allows for easy setup and operation remotely without special user software. Software updates can be accomplished remotely over the network connection.

The stations will be designed with a one-year service interval in mind. SOH monitoring and logging allows the system to remotely notify the user in the event of any out of range conditions.

Communications

The Phase I goals in this area were to investigate the following:

- Direct satellite network communications.
- Short haul wireless network communications.
- Methods of data compression and error recovery.

Satellite communications at this time consist mainly of two types; geosynchronous earth orbit (GEO) and LEO, as shown in Figure 4.

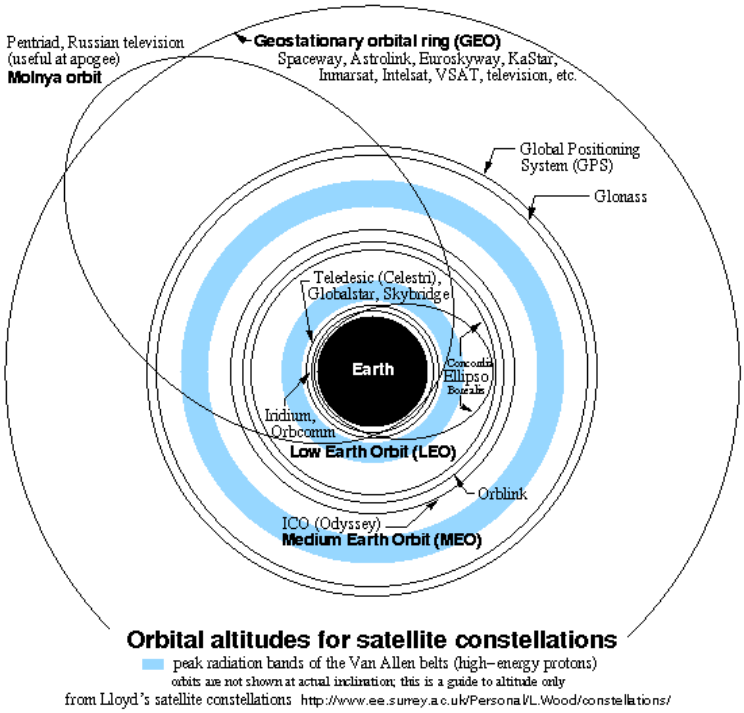


Figure 4. Satellite constellations and orbits.

GEO satellites can provide large coverage with a few satellites (but poor coverage in the polar regions) and can provide high data rates. However, they generally require a relatively large dish antenna, have large power consumption, and have large transmission delays to cope with. Examples of GEO systems are very small aperture terminal (VSAT) systems from various providers and Inmarsat.

Geotech currently provides VSAT solutions to its customers and has an on site VSAT link for testing. However, VSAT would not be the ideal solution for this project due to its large antenna and power requirements.

Inmarsat Broadband Global Area Network (BGAN) service is an interesting option for a GEO system since it only requires a small flat panel antenna (similar in size to a small laptop computer) making it much more portable and easy to setup. It can provide data rates up to 432 Kbps, but current data terminal modems require high power at around 50 W. However, as the BGAN service is rolled out for global coverage in 2005 and 2006, smaller lower power data terminals should become available.

LEO satellites provide the advantages of small omni-directional antennas, lower power operation, negligible transmission delays, and full earth coverage (including the polar regions). However, because of LEO, satellite positions in the sky are not fixed and many more satellites are required for full coverage. Since the satellites are constantly moving, complex satellite signal switching and routing are also required. This all makes the cost of a LEO system far more than a GEO system causing them to be less economically viable. Many LEO systems have been proposed but few have made it off the ground (Teledesic being the most ambitious to have failed and gone out of business). Of those that have, only Iridium and Globalstar have potential application to this project. Two systems, Skybridge and S2Com, are still on the drawing board and hold great promise to provide true "DSL Internet in the sky" if they are launched in the next few years.

Iridium began with great fanfare in the late 1990s, but quickly fell into bankruptcy. Iridium has been reborn and currently provides voice and data services globally. There are several low power (<2 W) original equipment manufacturer (OEM) data modems available for iridium, but data rates are limited to the 2400–9600 bps range.

Globalstar also provides voice and data service at 9600 bps. However, because it uses a simpler "bent pipe" architecture, its ground stations limit its coverage. Coverage is currently provided to North America, South America, Europe, Australia, and most of Asia. Low-power OEM data modems and small antennas are also available for the Globalstar system.

Geotech is currently fielding wireless local-area network (WiLAN) solutions for connecting stations to the network at data rates from 1.5 to 11 Mbps over distances of 20 miles.

With its SMART-24^{TM} instruments, Geotech has implemented and tested the following communications related technologies:

- TCP/IP based stack and transport providing error recovery and retransmission.
- CD 1.1 continuous data format providing error recovery, data buffering, and retransmission.
- Data compression providing up to 6:1 compression (low noise, 2-3:1 typical with high seismic background level).

The decision was made to not integrate the data modem directly into the digitizer package, but to rather provide standard ethernet and serial interfaces to a communication subsystem. Since at this time there is no universal communication solution to meet all needs, this provides the most flexible solution to allow any TCP/IP based solution to be used now and in the future as the need and technology presents itself.

In analyzing the communication requirements of a typical seismic data acquisition station with three channels running at 40 sps with authentication and compression turned on, the minimum required bit rate was found to be 9600 bps to account for data, commands, and protocol overhead at a ten second data frame interval. Larger data frame intervals do not appreciably increase transmission efficiency as shown in Table 2 and Figure 5. In Phase II of this project, Geotech will purchase and test various communication solutions that meet this requirement.

Table 2. Communication bandwidth requirements at various data frame intervals. This assumes three channels at 40 sps with 2:1 data compression and 50% overhead for command and protocol requirements.

Data Frame Size (sec)	No Compression (bps)	2:1 Compression (bps)
1	18048	14208
10	8716	4876
20	8198	4358

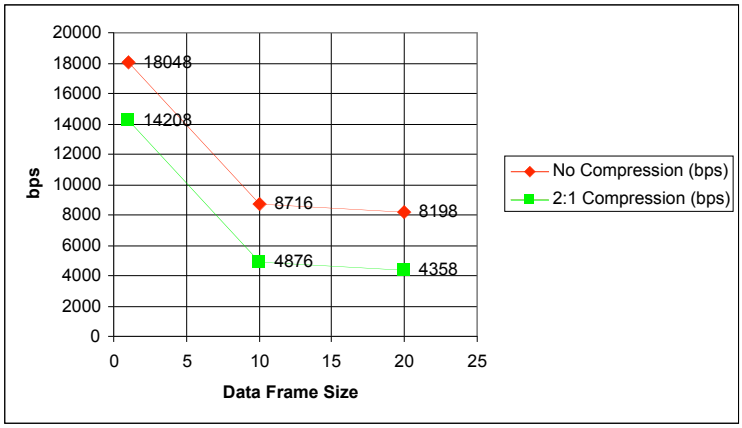


Figure 5. Communication bandwidth requirements at various data frame intervals. This assumes three channels at 40 sps with 2:1 data compression and 50% overhead for command and protocol requirements.

Advanced Power Supplies

The Phase I goal in this area was to investigate advanced power supply alternatives. Several fuel cell technologies are in the final stages of research and are developing into the first generation of fuel cell products. Four of these fuel cells technologies are direct methanol fuel cells (DMFCs), proton exchange membrane fuel cells (PEMFCs), direct ethanol fuel cells (DEFCs), and microbial fuel cells (MFCs). While each fuel cell technology has its own unique advantages, each also has disadvantages. MFCs are in a very infant stage and are not a viable option for the near future. While PEMFC has the highest energy density, the logistics of transporting and long-term storage of the hydrogen fuel eliminate PEMFC as a viable solution. In PEMFCs, converting H_2 energy to electric energy can be obtained at around 50%. If long-term storage and transportation issues are resolved in the next few years, this option could be revisited. DMFCs and DEFCs are the two promising options for near term implementation.

Figure 6. These are examples of 1st generation fuel cell products currently emerging onto the market.

The first generation of DMFC units will be released in 2005. Interest in DMFC (mainly due to the portability of the methanol fuel source) has accelerated the development of these units. Methanol can be transported similar to petroleum products. An estimated 10 gal. of methanol can run a 5-W station for one full year without refueling. This is roughly half the size of a fuel tank on a full sized American automobile. The weight of 10 gal. of methanol is approximately 66 lb, and that amount of methanol takes roughly 1.33 cu ft of space. This weighs less than and is much smaller than a solar panel. The only byproducts of a DMFC are water and CO_2.

Methanol does have some hazardous properties, while ethanol on the other hand does not. Ethanol has half the energy of methanol and would require a 20-gal. tank for storage of a year supply of ethanol. Ethanol is also transported as a liquid. A major disadvantage in using ethanol would be vandalism. Since ethanol is the consumable

portion of an alcoholic beverage, great care would need to be taken to secure the storage facility. DEFCs are considered an emerging fuel cell technology. Since ethanol is generated with renewable fuel sources, this is an interesting technology to consider even for the near future.

With pros and cons evaluated for each fuel cell technology, DMFC seems to be the most viable option for the near future. The following positives outweigh any of the negatives:

- 1st generation available in 2005.
- Fuel is readily available.
- 10 gal. (1.33 cu ft) needed for 1-year operation (66 lb).
- Byproducts are H_20 and CO_2

As DMFCs rapidly mature, they are slated to be the preferred fuel cell technology for small portable devices. At <5 W, a completely operational seismic station would be considered a small device. A typical laptop computer consumes 30–50 W. The first generation of DMFCs requires a lead-acid battery for startup.

With enough forethought, the installation can be adapted to support DMFCs in the initial implementation and DEFCs as the technology advances.

CONCLUSION AND RECOMMENDATION

With the convergence of the development of Geotech's new SMART-24™ instruments and this SBIR Phase I effort, Geotech has developed, implemented, and proven several technologies relating to this SBIR effort including the following:

- A 2nd generation 24-ADC with better performance and much lower power.
- Matching of sensors and ADC to improve performance.
- SMART-24™ technology for "plug and play" operation.
- TCP/IP connectivity making the instruments true "Internet appliances" with HTTP Web, Telnet, and FTP access.
- Standard data formats with CD 1.1 and error recovery.
- Data compression.
- Remote software updates via the Internet.
- Reliable VSAT satellite communications.

These advancements are already implemented in Geotech's new instruments and are in use by customers in the field.

Consumer mass market products (cell phones, PDAs, MP3 players, etc.) are propelling the push for smaller, low power components that will allow for further significant reductions in size and power of these instruments. At the same time there are now some viable options for low power satellite and wireless communications that were not previously available. New and faster communications options are also on the horizon and will become available in the next few years. Advanced power sources such as fuel cell technology are maturing with many products coming on line in the next year or so. For these reasons, Phase I of this project has shown that it is not only feasible, it is the ideal time to push forward and continue the development of a RRSS into Phase II.

OPTIMIZING DATA ACCESS AND AVAILABILITY FOR SEISMIC CALIBRATION RESEARCH

Michael L. Begnaud, Richard J. Stead, Julio Aguilar-Chang, and Hans E. Hartse

Los Alamos National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-36

ABSTRACT

The Ground-based Nuclear Explosion Monitoring Research & Engineering (GNEM R&E) program has made recent advances in optimizing data access and availability for seismic calibration research. Some of the most challenging tasks of maintaining functional and accessible data warehouses are the development of software to automate the continuous and up-to-date population of the database, the quality control (QC) needed to resolve data conflicts, the synchronization of database tables between unclassified and classified warehouses, and the integration of all data sources into a cohesive database for delivery to the Knowledge Base (KB).

One important challenge in using large data warehouses is the simple and efficient access to the vast holdings within them. Web-based tools have become important assets that address this problem. We have developed web-based tools that enable researchers to do tasks such as track the progress of seismic analysis, access information about stations, origins, and waveforms, view contextual information on a map, handle logistical tasks (i.e., assignment of unique identifiers, track the description and resolution of data problems identified through quality controls), and gain fast access to database metadata (e.g., schema descriptions).

Advances in easy access to metadata are supporting many of the higher-level efforts in quality control, automation, and web access. The first of these is the documentation of the seismic calibration schema using a database schema. This schema is designed to represent all of the detailed table and field information that, up until recently, has been available only in text-based documents. Such information in database form has immediate application to a wide variety of efforts involving the database. (e.g., table creation and quality control, software tools). Another advance in using metadata, with a more narrow application, has been the creation of bulletin descriptive tables. These tables describe the sources of bulletin data that have been imported into the data warehouse, as well as provide a means to track individual data elements to the corresponding lines of text in the original document.

As data become more voluminous and complex, QC has become an increasingly visible and important issue regarding the Knowledge Base. Improvements in QC procedures are helping researchers and data managers to more readily identify complex quality problems. The outcome is consistent research products resulting from improved data upon which those products are based. As we understand the QC problem in more detail, we have begun to automate the process of applying QC to large datasets.

Calibration efforts by Los Alamos National Laboratory (LANL) researchers require working with three separate data warehouses that are physically unable to communicate with each other: two are within LANL; one is located at a remote site. While it is relatively simple to add new data to all warehouses, it is difficult to capture changes made in one and then propagate them to the other two. We have recently developed a procedure based on database triggers to capture these changes. These triggers capture all update, insert, and delete operations against a predefined set of tables. Periodically, the information captured by these triggers is moved to the other environments and executed, thus keeping the warehouses synchronized.

OBJECTIVE(S)

The GNEM R&E program has made recent advances in applying data warehouses to seismic calibration research. Some of the most challenging tasks of maintaining functional data warehouses are the development of software to easily access the contents of the data warehouse, the QC needed to resolve data conflicts, the synchronization of database tables between local and remote warehouses, and the integration of all data sources into a cohesive database for delivery to the (KB). This paper is a brief introduction to the wide range of data management technical issues that we face everyday and the future work needed to fully address all aspects of managing and handling vast amounts of data in a data warehouse that is used in nuclear explosion monitoring research.

RESEARCH ACCOMPLISHED

Web Technology Access to Data Warehouses

As data gathering techniques continue to improve and general data availability increases, the GNEM R&E data warehouses will acquire more data than is readily accessible using the standard SQL command-line interface. One of the challenges is to develop a simple, yet efficient way to view the contents of our data warehouses to assist researchers in developing their calibration products.

Web-based tools provide an efficient, yet easy way to access data from the Oracle GNEM R&E databases. In addition to viewing the seismic data itself, we have developed web pages to interact with database schema viewing and development, handle logistical tasks (i.e., assignment of unique identifiers, tracking of database problems and data requests from researchers, etc.), and view metadata contents (e.g., glossary). The LANL GNEM R&E intranet web technology interface has been operational for over a year and has been extremely useful for accessing data quickly. Recent improvements include being able to generate an origin query, viewing quick, interactive maps of queried data, and implementing a data request system for researchers. Figure 1 is a view of the starting LANL GNEM R&E home page.

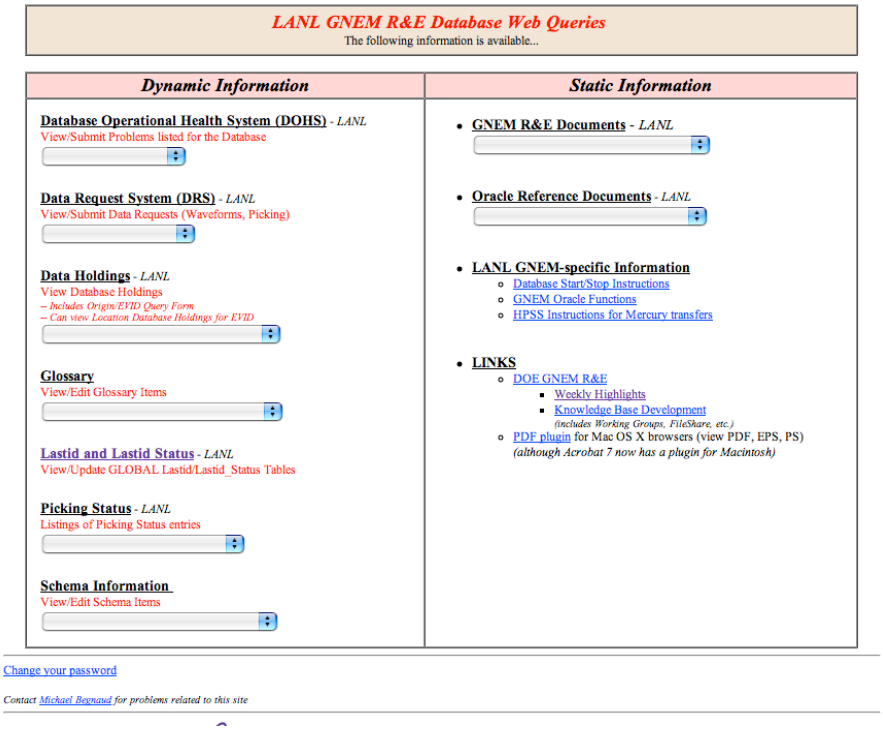


Figure 1. LANL GNEM starting web page. From this page, users can access seismic database entries, glossary and schema information, logistical data, and GNEM-related internal pages.

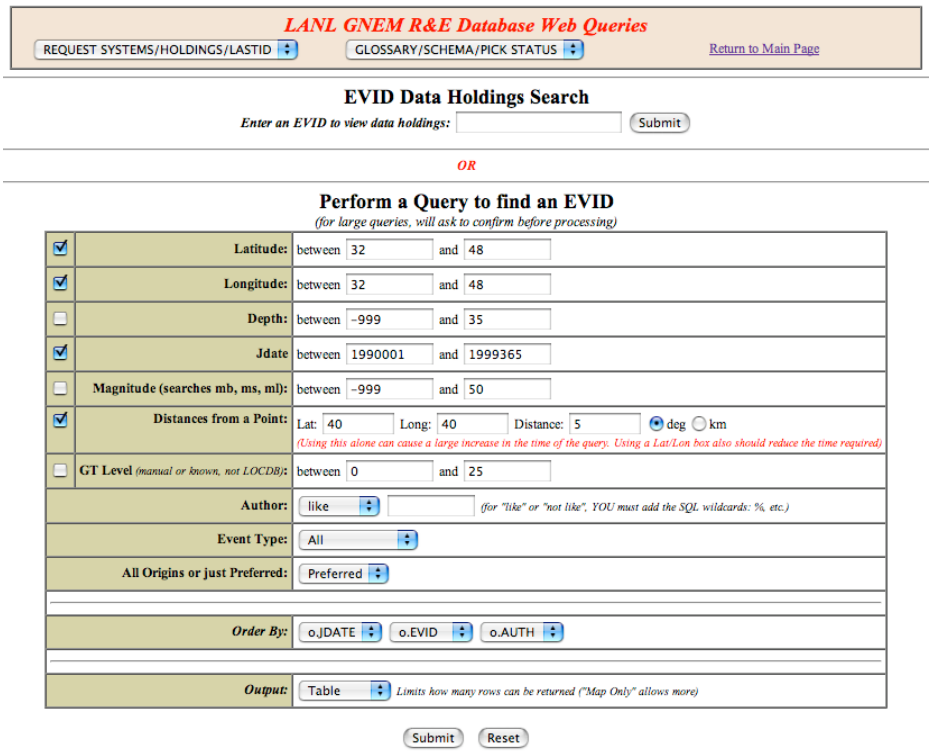


Figure 2. Event query page. Users can enter an EVID directly or perform an origin query. Selected parameters shown were used to generate a query producing results in Figure 3 and Figure 4.

Because of the need for LANL GNEM R&E team members to access data from remote locations, we implemented username/password access as well as Secure Socket Layer (SSL) 128-bit encryption for our internal web technology data access. Within LANL, users are granted access by a browser-standard basic username/password authentication. Outside LANL, users must first use a LANL-authorized cryptocard to gain web access behind the LANL firewall before proceeding to the username/password screen.

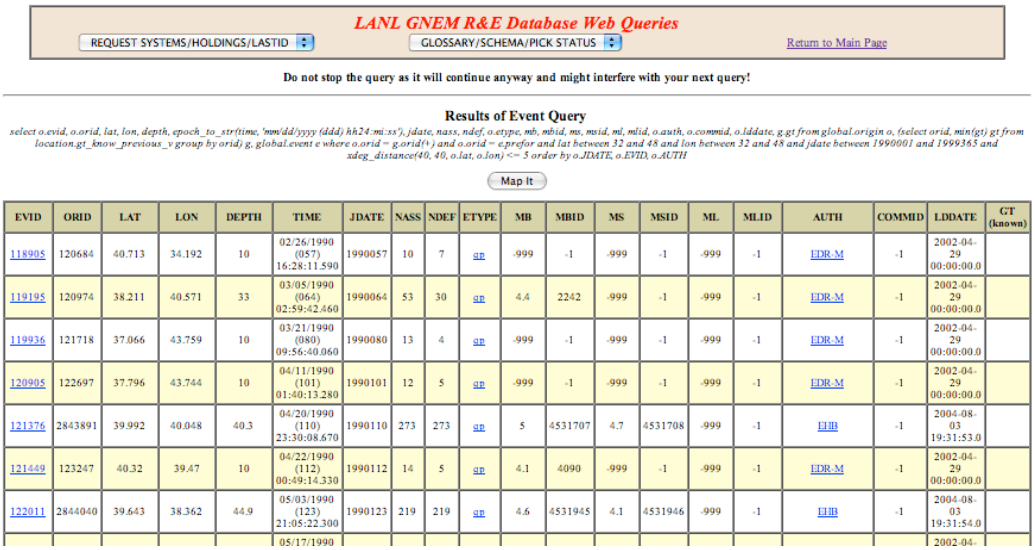
Seismic Data Holdings

The GNEM R&E database schema generally follows the National Nuclear Security Administration (NNSA) structure (Carr, 2005). This structure is mostly centered around events which are built with origins, associations, arrivals, waveforms, etc. Using an Event Identification number (EVID), a user can generally access all the data available for that event. Database users can either enter an EVID directly, or enter parameters for an origin query. Parameters include latitude/longitude, depth, julian date, magnitude, distance from a point, ground-truth value, author or authority, general event type (earthquake, explosion, mining), and all origins or just preferred (Figure 2). Users can request an output HTML table (Figure 3) or just gather data on the web server to produce a map (Figure 4).

In the table output view (Figure 3), a user will see origin information as well as the known ground-truth level. The AUTH and ETYPE fields have cross-referenced links to a glossary table. Clicking one of these links shows the definition of the field information. In the map output view (Figure 4), users can interactively view the results of the origin (or other) queries.

When an EVID is selected from the table output, a new screen appears with an initial summary of available data for that EVID (Figure 5). The web scripts determine if waveform segments, pick status entries, and location database entries exist for that event. Buttons are highlighted for those types available. Since there can be multiple origins for an event, the web page displays each origin and highlights buttons if arrivals, magnitudes, and amplitudes exist for that origin. Users can quickly navigate all data associated with an EVID.

In addition to the EVID-based data holdings, users can query for site-related information, by station abbreviation, reference station, or by entering a manual query (Figure 6). For a single site, the relevant SITE, AFFILIATION, SITECHAN, SENSOR, and INSTRUMENT information are displayed.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

Do not stop the query as it will continue anyway and might interfere with your next query!

Results of Event Query

select o.evid, o.orid, lat, lon, depth, epoch_to_str(time, 'mm/dd/yyyy (ddd) hh24:mi:ss'), jdate, nass, ndef, o.etype, mb, mbid, ms, msid, ml, mlid, o.auth, o.commid, o.lddate, g.gt from global.origin o, (select orid, min(gt) gt from location.gt_know_previous_v group by orid) g, global.event e where o.orid = g.orid(+) and o.orid = e.prefor and lat between 32 and 48 and lon between 32 and 48 and jdate between 1990001 and 1999365 and xdeg_distance(40, 40, o.lat, o.lon) <= 5 order by o.JDATE, o.EVID, o.AUTH

Map It

EVID	ORID	LAT	LON	DEPTH	TIME	JDATE	NASS	NDEF	ETYPE	MB	MBID	MS	MSID	ML	MLID	AUTH	COMMID	LDDATE	GT (known)
118905	120684	40.713	34.192	10	02/26/1990 (057) 16:28:11.590	1990057	10	7	gr	-.999	-1	-.999	-1	-.999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
119195	120974	38.211	40.571	33	03/05/1990 (064) 02:59:42.460	1990064	53	30	gr	4.4	2242	-.999	-1	-.999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
119936	121718	37.066	43.759	10	03/21/1990 (080) 09:56:40.060	1990080	13	4	gr	-.999	-1	-.999	-1	-.999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
120905	122697	37.796	43.744	10	04/11/1990 (101) 01:40:13.280	1990101	12	5	gr	-.999	-1	-.999	-1	-.999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
121376	2843891	39.992	40.048	40.3	04/20/1990 (110) 23:30:08.670	1990110	273	273	gr	5	4531707	4.7	4531708	-.999	-1	EHB	-1	2004-08-03 19:31:53.0	
121449	123247	40.32	39.47	10	04/22/1990 (112) 00:49:14.330	1990112	14	5	gr	4.1	4090	-.999	-1	-.999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
122011	2844040	39.643	38.362	44.9	05/03/1990 (123) 21:05:22.300	1990123	219	219	gr	4.6	4531945	4.1	4531946	-.999	-1	EHB	-1	2004-08-03 19:31:54.0	
					05/17/1990													2002-04-	

Figure 3. Partial results of table output for origin query using parameters from Figure 2. Users may select the EVID at left to view specific data (origins, arrivals, netmags, etc.) for that event. Other terms are cross-referenced with glossary tables, giving the definition of the term.

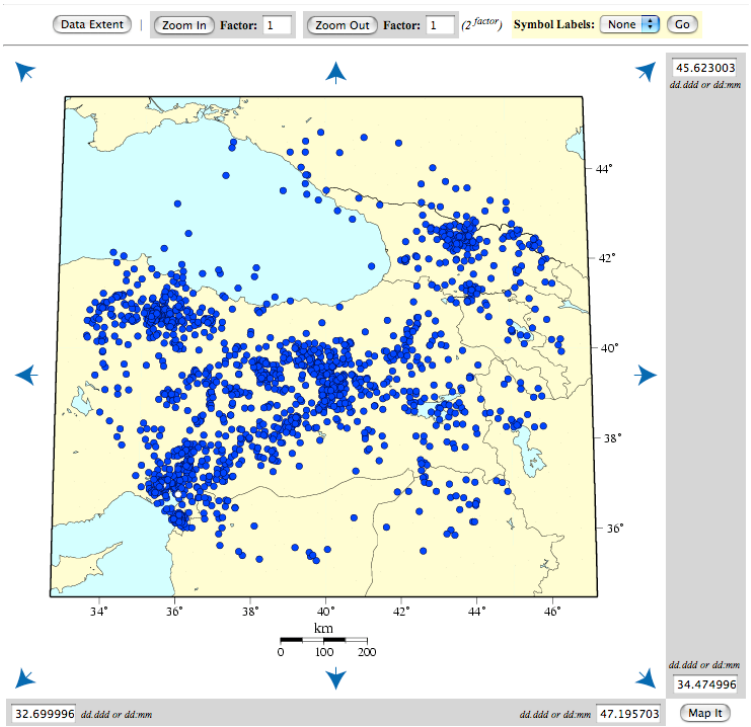
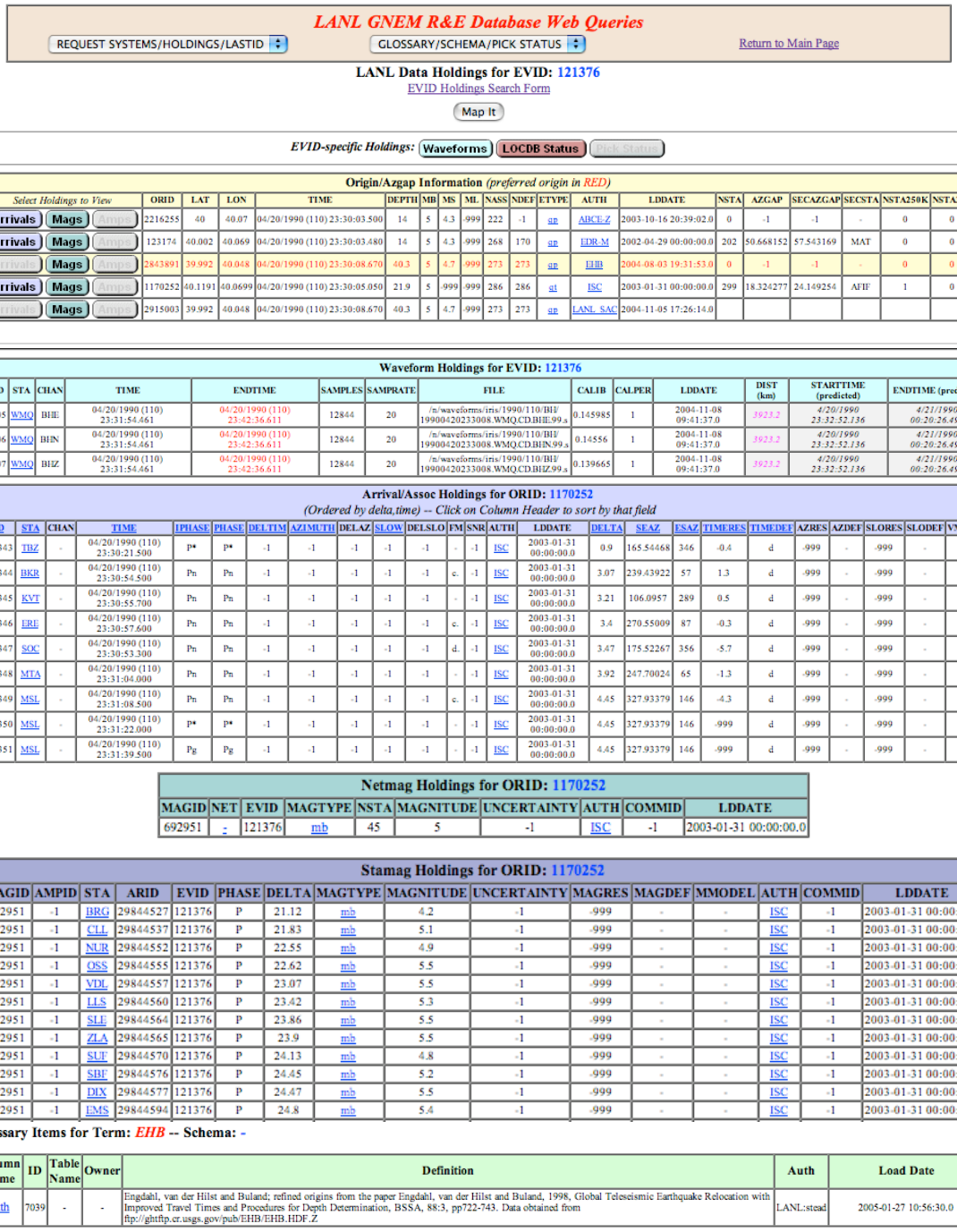


Figure 4. Interactive map of origins produced from query in Figure 2. Users can zoom in or out, set the bounds of the map, and translate the view. Labels can also be turned on or off.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

LANL Data Holdings for EVID: 121376

EVID Holdings Search Form

Map It

EVID-specific Holdings: Waveforms | LOCDB Status | Pick Status

Origin/Azgap Information *(preferred origin in RED)*

Select Holdings to View	ORID	LAT	LON	TIME	DEPTH	MB	MS	ML	NASS	NDEF	ETYPE	AUTH	LDDATE	NSTA	AZGAP	SECAZGAP	SECSTA	NSTA250K	NSTA30K
Arrivals Mags Ampls	2216255	40	40.07	04/20/1990 (110) 23:30:03.500	14	5	4.3	.999	222	-1	sp	ABCE-Z	2003-10-16 20:39:02.0	0	-1	-1	-	0	0
Arrivals Mags Ampls	123174	40.002	40.069	04/20/1990 (110) 23:30:03.480	14	5	4.3	.999	268	170	sp	EDR-M	2002-04-29 00:00:00.0	202	50.668152	57.543169	MAT	0	0
Arrivals Mags Ampls	2843891	39.992	40.048	04/20/1990 (110) 23:30:08.670	40.3	5	4.7	.999	273	273	sp	EHB	2004-08-03 19:31:53.0	0	-1	-1	-	0	0
Arrivals Mags Ampls	1170252	40.1191	40.0699	04/20/1990 (110) 23:30:05.050	21.9	5	.999	.999	286	286	at	ISC	2003-01-31 00:00:00.0	299	18.324277	24.149254	AFIF	1	0
Arrivals Mags Ampls	2915003	39.992	40.048	04/20/1990 (110) 23:30:08.670	40.3	5	4.7	.999	273	273	sp	LANL_SAC	2004-11-05 17:26:14.0						

Waveform Holdings for EVID: 121376

WFID	STA	CHAN	TIME	ENDTIME	SAMPLES	SAMPRATE	FILE	CALIB	CALPER	LDDATE	DIST (km)	STARTTIME (predicted)	ENDTIME (predicted)
348405	WMQ	BHE	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/19900420233008.WMQ.CD.BHE.99.s	0.145985	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498
348406	WMQ	BHN	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/19900420233008.WMQ.CD.BHN.99.s	0.14556	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498
348407	WMQ	BHZ	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/19900420233008.WMQ.CD.BHZ.99.s	0.139665	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498

Arrival/Assoc Holdings for ORID: 1170252

(Ordered by delta,time) -- Click on Column Header to sort by that field

ARID	STA	CHAN	TIME	IPHASE	PHASE	DELTIM	AZIMUTH	DELAZ	SLOW	DELSLO	FM	SNR	AUTH	LDDATE	DELTA	SEAZ	ESAZ	TIMERES	TIMEDEF	AZRES	AZDEF	SLORES	SLODEF	VMODEL
29844343	TBZ	-	04/20/1990 (110) 23:30:21.500	P*	P*	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	0.9	165.54468	346	-0.4	d	.999	-	.999	-	-
29844344	BKR	-	04/20/1990 (110) 23:30:54.500	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	3.07	239.43922	57	1.3	d	.999	-	.999	-	-
29844345	KVT	-	04/20/1990 (110) 23:30:55.700	Pn	Pn	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	3.21	106.0957	289	0.5	d	.999	-	.999	-	-
29844346	ERE	-	04/20/1990 (110) 23:30:57.600	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	3.4	270.55009	87	-0.3	d	.999	-	.999	-	-
29844347	SOC	-	04/20/1990 (110) 23:30:53.300	Pn	Pn	-1	-1	-1	-1	-1	d.	-1	ISC	2003-01-31 00:00:00.0	3.47	175.52267	356	-5.7	d	.999	-	.999	-	-
29844348	MTA	-	04/20/1990 (110) 23:31:04.000	Pn	Pn	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	3.92	247.70024	65	-1.3	d	.999	-	.999	-	-
29844349	MSL	-	04/20/1990 (110) 23:31:08.500	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	-4.3	d	.999	-	.999	-	-
29844350	MSL	-	04/20/1990 (110) 23:31:22.000	P*	P*	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	.999	d	.999	-	.999	-	-
29844351	MSL	-	04/20/1990 (110) 23:31:39.500	Pg	Pg	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	.999	d	.999	-	.999	-	-

Netmag Holdings for ORID: 1170252

MAGID	NET	EVID	MAGTYPE	NSTA	MAGNITUDE	UNCERTAINTY	AUTH	COMMID	LDDATE
692951	-	121376	mb	45	5	-1	ISC	-1	2003-01-31 00:00:00.0

Stamag Holdings for ORID: 1170252

MAGID	AMPID	STA	ARID	EVID	PHASE	DELTA	MAGTYPE	MAGNITUDE	UNCERTAINTY	MAGRES	MAGDEF	MMODEL	AUTH	COMMID	LDDATE
692951	-1	BRG	29844527	121376	P	21.12	mb	4.2	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	CLL	29844537	121376	P	21.83	mb	5.1	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	NUR	29844552	121376	P	22.55	mb	4.9	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	OSS	29844555	121376	P	22.62	mb	5.5	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	VDL	29844557	121376	P	23.07	mb	5.5	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	LLS	29844560	121376	P	23.42	mb	5.3	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SLE	29844564	121376	P	23.86	mb	5.5	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	ZLA	29844565	121376	P	23.9	mb	5.5	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SUF	29844570	121376	P	24.13	mb	4.8	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SBF	29844576	121376	P	24.45	mb	5.2	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	DIX	29844577	121376	P	24.47	mb	5.5	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	EMS	29844594	121376	P	24.8	mb	5.4	-1	.999	-	-	ISC	-1	2003-01-31 00:00:00.0

Glossary Items for Term: *EHB* -- Schema: -

Column Name	ID	Table Name	Owner	Definition	Auth	Load Date
auth	7039	-	-	Engdahl, van der Hilst and Buland; refined origins from the paper Engdahl, van der Hilst and Buland, 1998, Global Teleseismic Earthquake Relocation with Improved Travel Times and Procedures for Depth Determination, BSSA, 88:3, pp722-743. Data obtained from ftp://ghtftp.cr.usgs.gov/pub/EHB/EHB.HDF.Z	LANL:stead	2005-01-27 10:56:30.0

Figure 5. View of EVID-specific data holdings page and several data frames, including a glossary definition. Buttons are highlighted if data are available for that data type. From this page, arrivals, magnitudes (netmag, stamag), and amplitude data can be viewed for the different origins. Users can also view a map of the different origins and stations with waveforms.

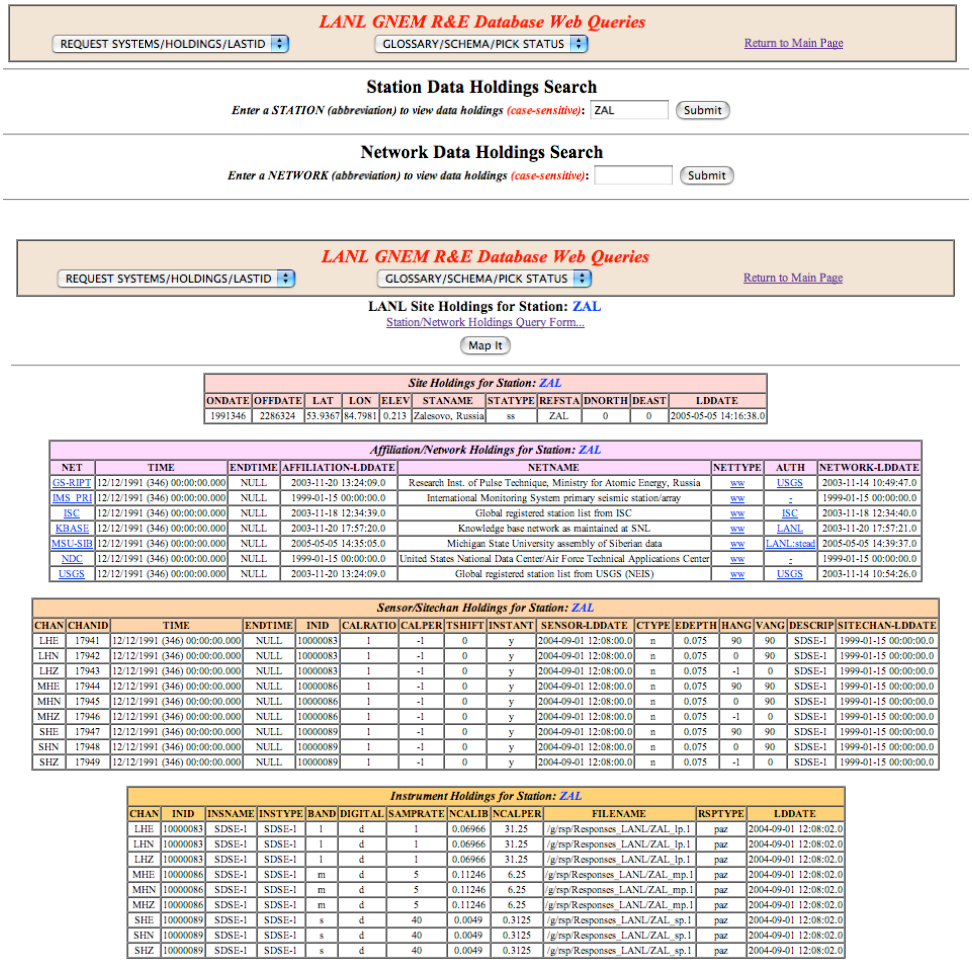
Schema Documentation

A major effort in making database metadata available is the documentation of the seismic calibration schema using a database schema. This schema is designed to represent all of the detailed table and field information that, up until recently, has been available only in text-based documents. These documents include versions of the NNSA KB core schema, NNSA KB custom schema, and the United States National Data Center (USNDC) schema documents. The portions that are most needed as readily available metadata are also the portions most amenable to adaptation into

database tables themselves: the table descriptions and the column description. Four tables are used to describe schema information: TABDESCRIPT, COLASSOC, COLDESCRIPT, and GLOSSARY. We have also developed a web technology interface to view and edit the schema and glossary information (Figure 7). The schema tables have been accepted as the schema documentation for the GNEM R&E program and are now used by Sandia National Laboratories for complete schema documentation (Carr, 2005). In addition, many of the KB tools developed by Sandia depend on these schema database tables.

TABDESCRIPT provides a basic description of the table, identifies that table with a particular documented schema, and provides a reference in the database to connect the fields that may be associated with the table.

The COLDESCRIPT table not only provides the description of a column but also provides metadata such as NA values, units, and ranges in useful forms. Most numeric ranges have been properly translated into *nmin*, *nminop*, *nmax*, and *nmaxop*. Each operation is relative to the value in the column; that is, 'column *nminop nmin'* and 'column *nmaxop nmax'*. If both are set, then both must apply (implied 'and'). A range type of 'defined' means the value of the column is limited to a short set of predefined values. A 'finite set' is a limited but long or not pre-defined set of values. A 'reference set' is limited to the values in a particular table (as given in *reftab*). Using these fields properly can completely and precisely define the various field ranges.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

Station Data Holdings Search

Enter a STATION (abbreviation) to view data holdings (case-sensitive): ZAL Submit

Network Data Holdings Search

Enter a NETWORK (abbreviation) to view data holdings (case-sensitive): Submit

LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

LANL Site Holdings for Station: ZAL

Station/Network Holdings Query Form...

Map It

Site Holdings for Station: ZAL

ONDATE	OFFDATE	LAT	LON	ELEV	STANAME	STATYPE	REFSTA	DNORTH	DEAST	LDDATE
1991346	2286324	53.9367	84.7981	0.213	Zalesovo, Russia	ss	ZAL	0	0	2005-05-05 14:16:38.0

Affiliation/Network Holdings for Station: ZAL

NET	TIME	ENDTIME	AFFILIATION-LDDATE	NETNAME	NETTYPE	AUTH	NETWORK-LDDATE
GS-RIPT	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 13:24:09.0	Research Inst. of Pulse Technique, Ministry for Atomic Energy, Russia	ww	USGS	2003-11-14 10:49:47.0
IMS_PRI	12/12/1991 (346) 00:00:00.000	NULL	1999-01-15 00:00:00.0	International Monitoring System primary seismic station/array	ww	-	1999-01-15 00:00:00.0
ISC	12/12/1991 (346) 00:00:00.000	NULL	2003-11-18 12:34:39.0	Global registered station list from ISC	ww	ISC	2003-11-18 12:34:40.0
KBASE	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 17:57:20.0	Knowledge base network as maintained at SNL	ww	LANL	2003-11-20 17:57:21.0
MSU-SIB	12/12/1991 (346) 00:00:00.000	NULL	2005-05-05 14:35:05.0	Michigan State University assembly of Siberian data	ww	LANL:stead	2005-05-05 14:39:37.0
NDC	12/12/1991 (346) 00:00:00.000	NULL	1999-01-15 00:00:00.0	United States National Data Center/Air Force Technical Applications Center	ww	-	1999-01-15 00:00:00.0
USGS	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 13:24:09.0	Global registered station list from USGS (NEIS)	ww	USGS	2003-11-14 10:54:26.0

Sensor/Sitechan Holdings for Station: ZAL

CHAN	CHANID	TIME	ENDTIME	INID	CALRATIO	CALPER	TSHIFT	INSTANT	SENSOR-LDDATE	CTYPE	EDEPTH	HANG	VANG	DESCRIP	SITECHAN-LDDATE
LHE	17941	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
LHN	17942	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
LHZ	17943	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0
MHE	17944	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
MHN	17945	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
MHZ	17946	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0
SHE	17947	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
SHN	17948	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
SHZ	17949	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0

Instrument Holdings for Station: ZAL

CHAN	INID	INSNAME	INSTYPE	BAND	DIGITAL	SAMPRATE	NCALIB	NCALPER	FILENAME	RSPTYPE	LDDATE
LHE	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
LHN	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
LHZ	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
MHE	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
MHN	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
MHZ	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
SHE	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0
SHN	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0
SHZ	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0

Figure 6. Site query and information pages for a single station (ZAL). In addition to standard site information, affiliation, sitechan, sensor, and instrument data are also displayed. Users may also plot a map of the site.

The COLASSOC table associates a given COLDESCRIPT with a TABDESCRIPT. The table will have multiple columns (column positions 1 through the total number, in order), and columns can appear in multiple tables. The column type will define the basic function of the field in the database (i.e., primary key, unique key, descriptive data, measurement data, administrative data). *Key* allows key columns to be identified with respect to the table: the reference table for the key, or a table in which the key is foreign. These keys are primarily numerical identifiers. Keyschema is used when the reference table is part of a separate schema.

The GLOSSARY table serves two purposes: The first is to simply define generic strings used in various description fields, primarily acronyms and abbreviations. The second is to serve as the reference table for 'defined' range types and 'finite set' range types. A given definition can apply in all circumstances (column_name, table_name, owner and schema are all not set), which is a generic definition, or it can apply to increasingly selective subsets of columns, tables, owners, and schemas. For 'defined' and 'reference set' range types in COLDESCRIPT, the complete permitted set of values will be found in GLOSSARY, one entry for each value (values are found in the name column), and *column_name* will always be set for these.

Having such schema information in database form has proven beneficial to a wide variety of efforts involving the database, both specific to Los Alamos, and across NNSA. This schema for table description metadata is a key contribution to the web-based database documentation discussed in this paper. It is also the foundation for automated QC efforts at LANL (see below). The schema tables have changed slightly over the past year, in response to experience using it at LANL and elsewhere. The tables provide immediate advantages in the maintenance of the schema descriptions, such that they can be easily checked for errors, quality, and completeness. The use of the TABDESCRIPT, COLDESCRIPT, and COLASSOC tables is becoming fairly well established. The GLOSSARY table has more recently seen greater use and has proven helpful in QC of text fields and in finding full descriptions of various text-based values such as phase names or authors (e.g., "just what is SPdifKS?", "is there a reference for "SIB:AT62"?").

Logistical Information

The LANL web technology access not only displays seismic and schema information, but has interfaces to allow researchers to handle logistical tasks such as requesting waveform, picking, and catalog data, submitting problems encountered with the database, and viewing identifier values for database primary key fields. Both of these functions

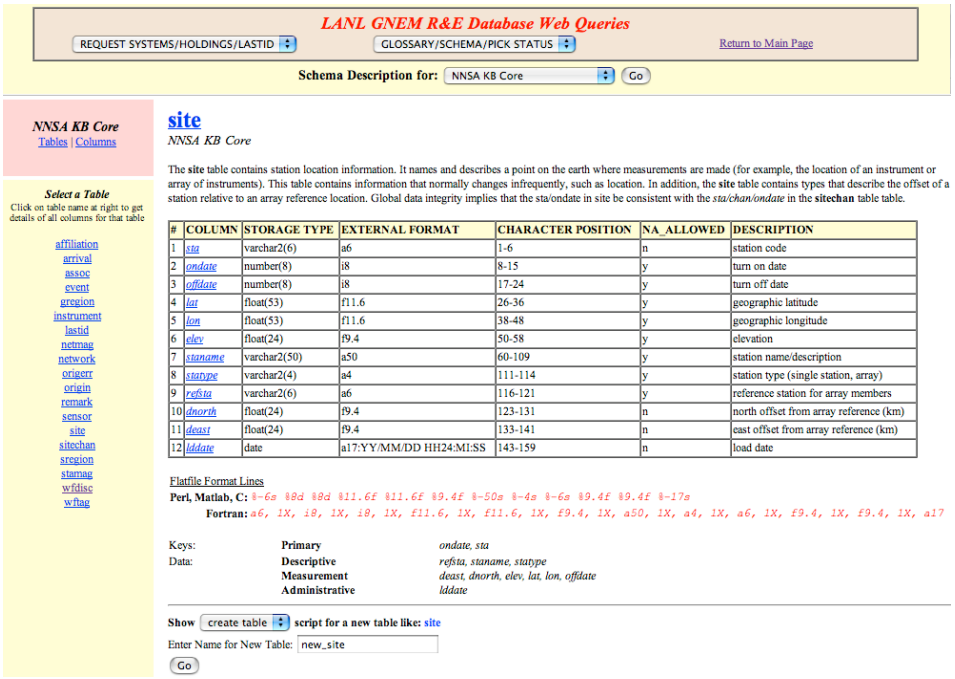


Figure 7. View of schema web page. Any schema can be displayed with links pointing to table descriptions as well as individual column descriptions. Other web pages are also used to directly edit the schema information.

enable tracking of pending and completed data requests and listed database problems.

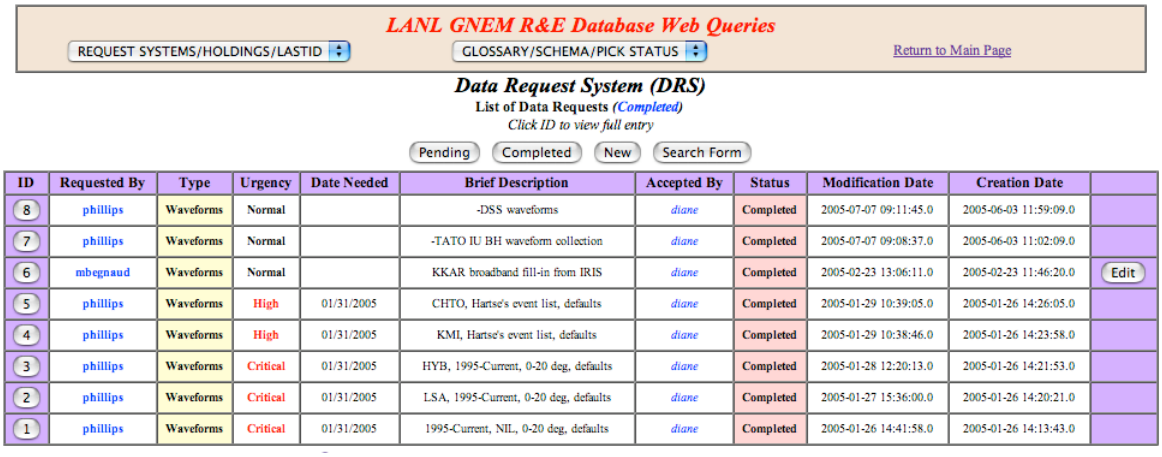
For the new Data Request System (DRS) (Figure 8), users can request that waveform, arrival picks, or catalog data be acquired for use in their research projects. Requests are entered into the system, logging who made the request along with the request details, and notifications are automatically sent to members of the LANL Data Management Team. A member of the team then accepts the request, retrieves the data, and sets the status of the request to "Completed." During this process, the person who originally made the data request is automatically notified of changes in status or comments from the "Acceptor." This formal method for data requests allows tracking of researcher needs and reduces instances of miscommunication.

Quality Control and Assurance

As data become more voluminous and complex, QC has become a challenging and extremely interesting issue to consider when developing content for the KB. In particular, because of recent advances in tools that access KB data, researchers have been able to make more efficient use of this large volume of data but have also found inconsistencies in applying the data to their research efforts. Improvements in QC procedures are helping researchers and data managers to more readily identify complex quality problems. The outcome is improved research products resulting from improved data upon which those products are based. QC is handled in a wide variety of ways at the present, and much effort is being made to better structure this procedure and automate as much of it as is practical to do so.

There are three main categories of QC: manual, tool-assisted, and automated. Improvements in manual QC are occurring constantly as a wider variety of issues are captured and understood. But this kind of QC is the least transportable and repeatable. The next step is to capture the tracking and resolution of QC problems in various simple tools, usually case-specific scripts. Tool-assisted QC is more transportable and repeatable, since the script serves as documentation of procedures but it still requires case-by-case modification and application. Automated QC is preferred and cannot happen without there first being a fairly comprehensive understanding of the problem.

The best approach to automated QC is to document exactly what the database should be and reject anything that does not conform. This approach is why the database descriptive schema discussed above has been a valuable tool in automated QC. This cannot address all QC issues (for example, QC of waveforms), but will handle the bulk of the information in the database. LANL now has an automated process that can be configured to run a very comprehensive QC against a wide variety of data sets that may be incorporated into the KB. Since it is based on the content of the schema tables (i.e., TABDESCRIPT, COLASSOC, COLDESCRIPT, GLOSSARY), it can readily handle the addition of new custom tables for particular data sets. It produces a comprehensive QC report that greatly speeds the identification of problems that need to be addressed. This was of great help, for example, in preparing the recently delivered Siberian dataset, which was a highly heterogeneous collection of data from a wide variety of sources. The process is based on the schema tables, uses a simple parameter file, and implements single-column and single-table tests, two-table joins, and the notorious "wftag"-type join. It also extends these tests and joins using a special database table called COMPLEXJOIN, that permits a wide variety of complex relationships including



ID	Requested By	Type	Urgency	Date Needed	Brief Description	Accepted By	Status	Modification Date	Creation Date	
8	phillips	Waveforms	Normal		-DSS waveforms	diane	Completed	2005-07-07 09:11:45.0	2005-06-03 11:59:09.0	
7	phillips	Waveforms	Normal		-TATO IU BH waveform collection	diane	Completed	2005-07-07 09:08:37.0	2005-06-03 11:02:09.0	
6	mbegnaud	Waveforms	Normal		KKAR broadband fill-in from IRIS	diane	Completed	2005-02-23 13:06:11.0	2005-02-23 11:46:20.0	Edit
5	phillips	Waveforms	High	01/31/2005	CHTO, Hartse's event list, defaults	diane	Completed	2005-01-29 10:39:05.0	2005-01-26 14:26:05.0	
4	phillips	Waveforms	High	01/31/2005	KMI, Hartse's event list, defaults	diane	Completed	2005-01-29 10:38:46.0	2005-01-26 14:23:58.0	
3	phillips	Waveforms	Critical	01/31/2005	HYB, 1995-Current, 0-20 deg, defaults	diane	Completed	2005-01-28 12:20:13.0	2005-01-26 14:21:53.0	
2	phillips	Waveforms	Critical	01/31/2005	LSA, 1995-Current, 0-20 deg, defaults	diane	Completed	2005-01-27 15:36:00.0	2005-01-26 14:20:21.0	
1	phillips	Waveforms	Critical	01/31/2005	1995-Current, NIL, 0-20 deg, defaults	diane	Completed	2005-01-26 14:41:58.0	2005-01-26 14:13:43.0	

Figure 8. Data Request System page showing "Completed" requests.

grouping relationships (i.e., comparing origin.nass to assoc), multi-table joins (ex: comparing wfdisc.instype to instrument.instype), and computations (i.e., comparing origin.time to origin.jdate). Future development in this area will examine the possibility of automating the repairs, in addition to just the QC. In general, QC remains a large problem and GNEM R&E is making new and unique contributions toward resolving this important problem.

Bulletin Descriptive Tables

An advance in using schema information for QC has been the creation of bulletin descriptive tables. These tables describe the sources of bulletin data that have been imported into the data warehouse, as well as providing a means to track individual data elements to the corresponding lines of text in the original document. There are two tables involved: BULLETIN and BULLASSOC. The BULLETIN table contains one entry for each individual bulletin, with columns *dir* and *dfile* pointing to the text file on the system that contains the bulletin. It also has a *bullid* column that is unique to each bulletin. The other information in the table describes the bulletin, including format. The BULLASSOC table has one line for each data object extracted from the bulletin (origins, arrivals, magnitudes, etc.) It links the *bullid* and the *id* of the extracted object and provides the line number in the bulletin corresponding to the object. These tables have immediate use in QC. First, they allow problematic objects to be traced directly to the corresponding file and line number. Second, they can be used to extract the entire contents of a single bulletin from the integrated database when the need to remove or replace the data from a particular bulletin arises.

Segmenting Continuous Waveforms

Over the years LANL has acquired segmented and continuous waveforms from many different sources in formats such as SEED, SAC, CSS (with accompanying WFDISC lines), GSE, and SEGY. To make these data readily available to researchers, we have developed a PERL code that uses a database interface (the “PERL DBI”) to assemble user-specified, event-based wave segments into SAC files. We call this code "wfdisc2sac.pl", because it requires a database WFDISC line description of each waveform that might be cut and transformed into SAC format.

For the case of SEED data handling, wfdisc2sac.pl calls the executable “rdseed”. To run the code, a user builds a list of EVIDs that correspond to events of interest and specifies a list of desired stations and channels. The user can also specify desired time window lengths of the final SAC waves based on Jeffreys-Bullen travel-time tables. If ORIGIN, SITE, and ARRIVAL tables are available, wfdisc2sac.pl will use the PERL DBI to query these tables and find information to populate the newly created SAC header fields. To prevent accidental recutting of segments already listed in the LANL WFDISC table, an option is available to check for existing segments prior to attempting a fresh cut on continuous data. A second PERL code builds WFDISC flat file lines that can be immediately inserted into the WFDISC table.

Database Synchronization - Capturing and Propagating Data Changes

Calibration efforts by LANL researchers require the use of three separate databases that are physically unable to communicate with each other: two within LANL, and one at a remote site. Because of the lack of direct communication between these databases, maintaining data synchronization between them is difficult. The content of these databases is such that some data are common to all the databases, some data are common between only two of the databases, while some data are allowed to exist only at the remote location. While it is relatively simple to add new data to all databases, it is difficult to capture changes such as updates or deletes made in one and then propagate them to the other two.

We have recently developed a procedure based on database triggers to capture changes made to core database tables. This procedure has been in place for about one year, and results to date have been satisfactory. The changes being monitored on the predefined set of tables are data inserts, updates, and deletes. This process is referred to as Capture Data Changes (CDC). The acronym, as well as the fundamental idea, is similar to Oracle’s Change Data Capture method of implementing the incremental recording of data changes. The main difference between Oracle’s implementation and LANL’s is that our method does not depend on a particular version of the Oracle Relational Database Management System (RDBMS). Oracle’s CDC method is directly tied to a specific application that must be installed, configured, and run against an Oracle 9i database. Our procedure is entirely based on database triggers, which are available on any version of the Oracle RDBMS; thus, our implementation is not tied to any particular version of the Oracle database and can be implemented on any platform.

The concept is simple. Database triggers are created against a predefined set of tables to be monitored. These triggers fire upon insert, update, or delete operations against these tables. The triggers capture unique information about a row being inserted, deleted, or updated. Our implementation of capturing only the information needed to uniquely identify a changed row in a table leads to significant disk space savings.

An interesting by-product of our synchronization procedure is that we not only capture unique information about the rows being modified in the tables being monitored, but we also capture the modification date and the name of the database user who made the modification to each row. This information, together with the built-in database auditing capabilities that can be enabled at the table level, can serve the secondary purpose of providing a security audit trail to be able to answer questions regarding changes made to critical production data.

The synchronization operation between the source database and the two target databases is a manual process at this time. The synchronization operation starts after a predetermined number of changes have occurred in the source database tables. A special set of tables is created from the unique information captured by the triggers that contain the table structure and changed rows of data from the source tables that will be used to replace the outdated information in the target databases. The second and final step of the synchronization process is done on the target databases, both at LANL and at the remote location. First, we delete from and insert rows into the target tables. The rows to be deleted in the target databases are the rows that were either updated or deleted from the original source tables. In this step, the decision was made to replace the entire row when an update occurred in the source table, rather than try to make a column-by-column comparison to only update specific columns in the target table. The latter choice would be costlier in terms of computer resources.

CONCLUSIONS AND RECOMMENDATIONS

Developing web technology interfaces to handle common database queries has allowed LANL researchers to access data more readily and efficiently. Specific information for seismic events can be retrieved quickly with ties to relevant information. Utilizing protected web interfaces allows users to view data remotely and helps in synchronizing data retrieval and processing tasks. We are continually finding new reasons and ways to securely access the database through the web. Future web technology development includes tracking processing steps for waveforms, picking, amplitudes, etc., as well as improving QC checks. In addition, we are working on better methods for viewing and interacting with waveform and map data via a web interface.

QC has become an increasingly visible and important issue regarding the KB, as data have become more voluminous and complex. Improvements in QC procedures will help researchers and data managers to more readily identify complex quality problems. The outcome is improved research products resulting from improved data upon which those products are based.

The process of CDC shows great promise. It is expected that we have not encountered all possible use cases, and modifications to the process will need to be made and perhaps auxiliary tables or triggers will need to be created. This process has been in place for about one year and it appears to be successful. Many of the steps involved in synchronizing local and remote databases are manual, and direct interaction with the databases is needed throughout the process. Future work includes the automation of many of the synchronization steps and the incorporation of appropriate quality control checks to ensure that the synchronization between databases was successful.

ACKNOWLEDGEMENTS

The authors wish to acknowledge LANL personnel who have made important contributions in the past and those who continue to make vital contributions to the development and maintenance of the LANL research data warehouse: Diane F. Baker, Marian D. Peters, W. Scott Phillips, George E. Randall, James T. Rutledge, and Steven R. Taylor,

REFERENCES

Carr, D. (2005), *National Nuclear Security Administration Knowledge Base Core Table Schema Document*, Sandia National Laboratories report SAND2002-3055.

NEW GROUND TRUTH CAPABILITY FROM INSAR TIME SERIES ANALYSIS

Sean M. Buckley[1], Paul Vincent[2], and Dochul Yang[1]

University of Texas at Austin[1] and Lawrence Livermore National Laboratory[2]

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract Nos. DE-FC52-03NA99566[1] and W-7405-ENG-48[2]

ABSTRACT

We demonstrate that next-generation interferometric synthetic aperture radar (InSAR) processing techniques applied to existing data provide rich InSAR ground truth content for exploitation in seismic source identification. InSAR time series analyses utilize tens of interferograms and can be implemented in different ways. In one such approach, conventional InSAR displacement maps are inverted in a final post-processing step. Alternatively, computationally intensive data reduction can be performed with specialized InSAR processing algorithms. The typical final result of these approaches is a synthesized set of cumulative displacement maps.

Examples from our recent work demonstrate that these InSAR processing techniques can provide appealing new ground truth capabilities. We construct movies showing the areal and temporal evolution of deformation associated with previous nuclear tests. In other analyses, we extract time histories of centimeter-scale surface displacement associated with tunneling. The potential exists to identify millimeter per year surface movements when sufficient data exists for InSAR techniques to isolate and remove phase signatures associated with digital elevation model errors and the atmosphere.

OBJECTIVES

The goal of this study is to assess new satellite SAR data and InSAR processing techniques for improvements to current remotely-sensed ground truth collection capability for determining seismic source location, depth and characterization. Specifically, we are

- applying InSAR time series analysis to sets of European Remote Sensing Satellite (ERS) and Envisat radar interferograms and
- using InSAR results to constrain geophysical source modeling to improve determinations of seismic source location, depth, and mechanism.

RESEARCH ACCOMPLISHED

This paper presents results from the application of an InSAR time series technique in which a linear inversion is applied to a database of interferograms to produce a sequence of interferograms which show cumulative deformation relative to a reference start date. We present results for two sites of interest: Lop Nor, China, nuclear test site and London, UK, underground tunnels. In both cases, we detect a time evolution of displacement signals that would not be possible using standard processing techniques. The new time series allow for more detailed analyses of the location and time history of subtle deformation signals and provide more accurate ground truth information from the same InSAR dataset.

Methodology

The goal is to obtain deformation measurements at each of the SAR acquisition dates for a set of interferograms overlapping in time and exhibiting generally good coherence. The result is a history of cumulative deformation relative to a user-selected SAR reference date.

This InSAR time series problem can be formulated as a linear system of equations (Usai, 2001; Schmidt and Bürgmann, 2003):

$$Ax = b\,, \qquad (1)$$

where b is an array of input interferogram displacements, x is the estimated cumulative deformation at each of the SAR acquisition dates, and A is a matrix relating the measurements to the parameters to be estimated. A single output date is associated with a column in A. The rows correspond to the input interferograms with the measured displacement assumed to be the difference between the estimated cumulative subsidence values between two image dates. Specifically, each row of A contains the following integer values: +1 at the interferogram reference date, –1 at the interferogram secondary date, and zeros for all other dates.

We find the solution x to the linear system through singular value decomposition. To address error propagation, error estimates associated with each of the input interferograms can be computed by summing the phase noise and Digital elevation model (DEM) error variances. Errors of the estimated displacements x are then computed as the square root of the main diagonal elements of the output covariance matrix:

$$\operatorname{cov} x = \left(V\Lambda^{-1}U^T\right)\operatorname{cov} b\left(V\Lambda^{-1}U^T\right)^T \qquad (2)$$

where $A = U\Lambda V$ is the singular value decomposition. The decomposition is readily obtained from any number of computational software packages.

The InSAR time series analysis examples presented in the next two sections were performed as an InSAR post-processing procedure. In other words, conventional differential InSAR processing was followed by an inversion of the geocoded displacement maps. Some other approaches, e.g., permanent scatterer analysis, require specialized intermediate processing steps to generate the final time series. Our post-processing strategy is motivated by a desire

to make best use of any number of conventional InSAR software packages and provide a means to easily extend the analysis of a study site when additional data is obtained.

Lop Nor, China Nuclear Test Site

InSAR data processing began with image formation for 15 SAR data acquisitions spanning 1996-1999 (Table 1). One hundred and five initial interferograms were created from the possible image-pair combinations. The maximum spatial and temporal baselines were 327 meters and 1191 days. The initial interferograms were evaluated and several discarded due to decorrelation. A mosaic of Shuttle Radar Topography Mission (SRTM) 3-arcsecond digital elevation model data was used to remove the topographic phase variations from the remaining interferograms.

The InSAR processing resulted in 101 geocoded maps consisting of radar line-of-sight displacements at 3-arcsecond horizontal postings. Errors corresponding to each output product pixel were assumed to come from two sources: phase noise associated with interferogram decorrelation and DEM errors. Phase standard deviations were calculated from the correlation coefficient, and a 10-meter SRTM one-standard deviation height error was used for time series error analysis. Atmospheric phase variations are also superimposed on the displacement measurements.

The 101 geocoded displacement and corresponding error maps were inverted to arrive at a time series of 14 synthetic displacement maps, one at each of the radar acquisition dates (Table 1, Figure 1). The 19960730 and 19980804 images are noticeably afflicted with atmospheric artifacts. Time series plots of several features of interest were extracted (Figure 2) and show ground motion both toward and away from the radar.

Figure 1 shows a time series of cumulative deformation observed of the Lop Nor, China, nuclear test site relative to a start date of January 1, 1996. Each frame represents cumulative deformation from the reference date (1/1/96) to the date labeled in the frame (e.g., the first frame labeled 960416 represents the cumulative deformation between Jan. 1, 1996 and April 16, 1996). The top row of frames shows no significant deformation occurring between 1/1/96 and 5/21/96. A deformation signal is suspected in the first frame of the second row (960730), but this frame is also contaminated with significant atmospheric noise. In the second frame of the second row, there is much less atmospheric noise, and the suspected deformation signal remains. This demonstrates how the time series helps identify and confirm suspect deformation signals. We know from published literature that there was an underground nuclear test conducted in the vicinity of the InSAR signal on 96/06/08 (Waldhauser et al., 2004). We are comparing these signals with those from the Nevada Test Site (Vincent et al., 2003) to learn more about the effect of rock type, etc. on the surface deformation signals from underground nuclear tests to be in a better position to use InSAR to help identify potentially clandestine nuclear test activities.

London, UK Underground Tunnels

We began our analysis of London with 31 SAR acquisitions distributed over the period 1992-2001 (Table 1). Due to the carefully controlled nature of the ERS-1 and ERS-2 orbits over Europe, we found all 465 potential SAR data pairs to have spatial baselines suitable for interferometry. We were able to visually inspect and remove initial interferograms with significant atmospheric variability associated with the 19960816, 19970801, 19971219, and 19990319 SAR acquisitions. SRTM DEM data was used to remove the topographic phase signature from interferograms with spatial baselines less than 200 meters. After performing consistency checks between combinations of interferograms, 211 3-arcsecond radar line-of-sight displacement maps were inverted using the InSAR time series approach previously discussed.

Figure 3 shows our time series results for London, UK, where a subsidence signal increases linearly with time over an underground utility tunnel owned by London Electric Company. Figure 4 shows a plot of the deformation time series associated with a single pixel in the deformation maps. A linear trend of indicates ongoing subsidence from 1993 through 2001. This tunnel is known to be 60 feet underground in a thick clay layer, which suggests that a soil creep dynamic may be responsible for the observed linear subsidence with time. We are currently working on a soil creep finite element model to simulate this creep process. The ability to identify the rate of subsidence with time is critical for accurate simulations of the subsurface processes causing the observed signals. We are also working on modeling similar tunnels in solid and fractured rock that have nonlinear deformation with time, as well as damage zones above past underground nuclear tests, e.g., Vincent et al. (2003).

CONCLUSIONS

InSAR time series approaches provide an organizational framework for the use and interpretation of InSAR deformation products resulting from anything more than a handful of SAR images. We have applied one such approach to two sites of interest: Lop Nor, China, and London, UK. Our results show detailed temporal and spatial information not easily obtained from the tens of displacements maps generated with conventional InSAR processing. The Lop Nor time series shows surface deformation appearing at the time of a known underground nuclear test as well as other interesting features that are under further investigation. The London time series shows the appearance and steady increase of subsidence with time associated with a known underground tunnel. The new observations afforded by InSAR time series analysis provide more detailed ground truth information which can be used to produce more accurate, time-dependent deformation models and better constrain seismic and aseismic subsurface sources of deformation. This information can ultimately help location and identification studies if a sufficient number (>10) of SAR image acquisitions are available.

ACKNOWLEDGEMENTS

We thank Krishnavikas Gudipati at UT-Austin for developing the InSAR linear inversion software used in this analysis. ERS original SAR data were provided by the European Space Agency copyright 1993-2000. Additional data purchases were made by the WInSAR consortium with funding from the National Aeronautics and Space Administration, United States Geological Survey, and the National Science Foundation.

REFERENCES

Schmidt, D. A. and R. Bürgmann (2003), Time-dependent land uplift and subsidence in the Santa Clara valley, California, from a large interferometric synthetic aperture radar data set, *J. Geophys. Res.* 108: (B9), 2416, doi:10:1029/2002JB002267.

Usai, S. (2001), A New Approach for Long Term Monitoring of Deformations by Differential SAR Interferometry, Ph.D. Dissertation, Delft University of Technology, The Netherlands.

Vincent, P., S. Larsen, D. Galloway, R. Laczniak., W. Walter, B. Foxall, and J. Zucca (2003), New signatures of underground nuclear tests revealed by satellite radar interferometry., *Geophy. Res. Let.,* 30: 2141.

Waldhauser, D., D. Schaff, P. G. Richards, W. Y. Kim (2004), Lop Nor revisited: underground nuclear explosion locations, 1976-1996, from double-difference analysis of regional and teleseismic data, *Bull Seim. Soc. Am*. 94, 1879-1889.

Table 1. Dates of SAR acquisitions used in Lop Nor and London InSAR time series analysis. Four London SAR acquisitions were discarded because they contained noticeable atmospheric artifacts.

Lop Nor

Date	Days Since Reference Date
19960101	0000
19960102	0001
19960205	0035
19960416	0106
19960520	0140
19960521	0141
19960730	0211
19970401	0456
19970819	0596
19971202	0701
19980106	0736
19980421	0841
19980804	0946
19980908	0981
19990406	1191

London

Date	Days Since Reference Date
19920505	0000
19920609	0035
19920818	0105
19920922	0140
19930209	0280
19930803	0455
19950413	1073
19950727	1178
19950831	1213
19950901	1214
19960816	Not used
19960920	1599
19970103	1704
19970314	1774
19970523	1844
19970627	1879
19970801	Not used
19971009	1983
19971114	2019
19971219	Not used
19980123	2089
19980403	2159
19990319	Not used
19990702	2614
19991015	2719
19991224	2789
20000127	2823
20000128	2824
20000407	2894
20001103	3104
20010112	3174

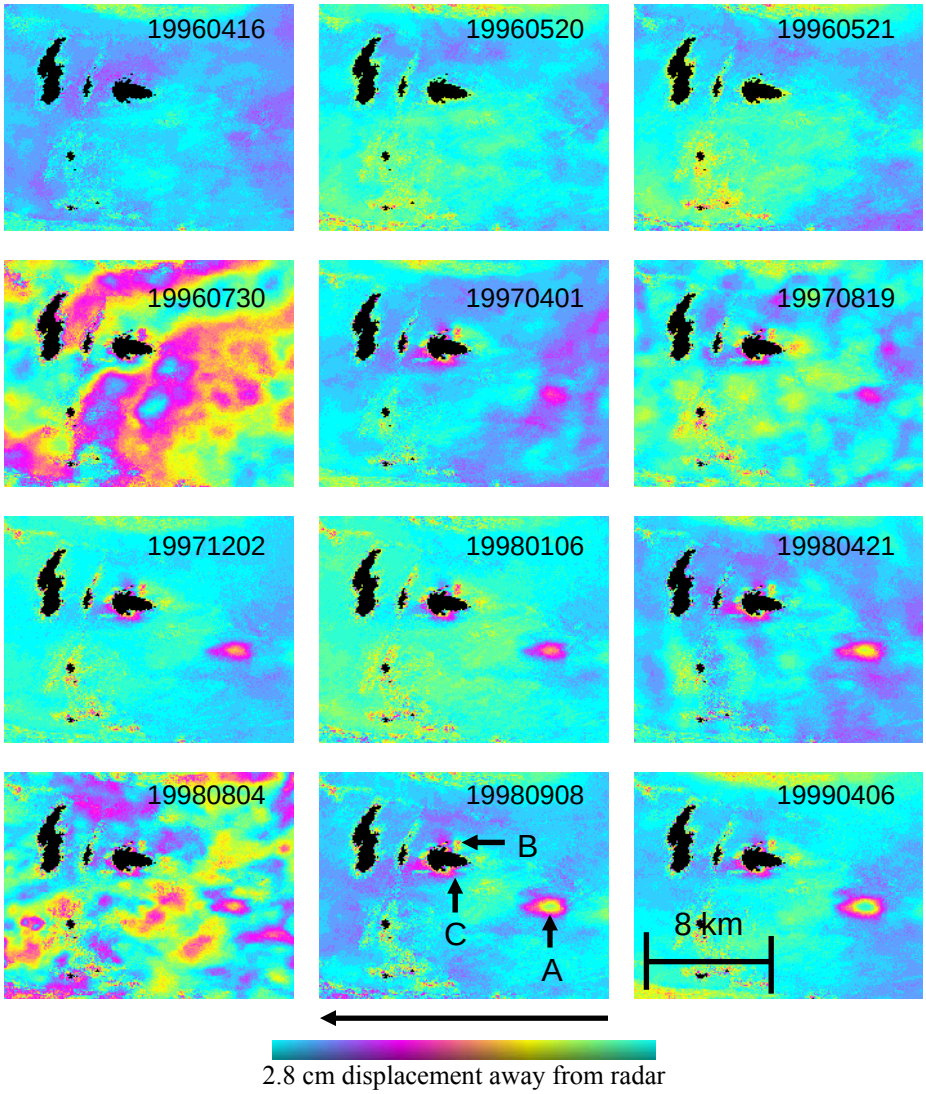


Figure 1. Maps of cumulative displacement since January 1, 1996 for Lop Nor. One cycle through color wheel represents 2.8 cm of radar line-of-sight displacement. The 19960730 and 19980804 images are noticeably afflicted with atmospheric artifacts.

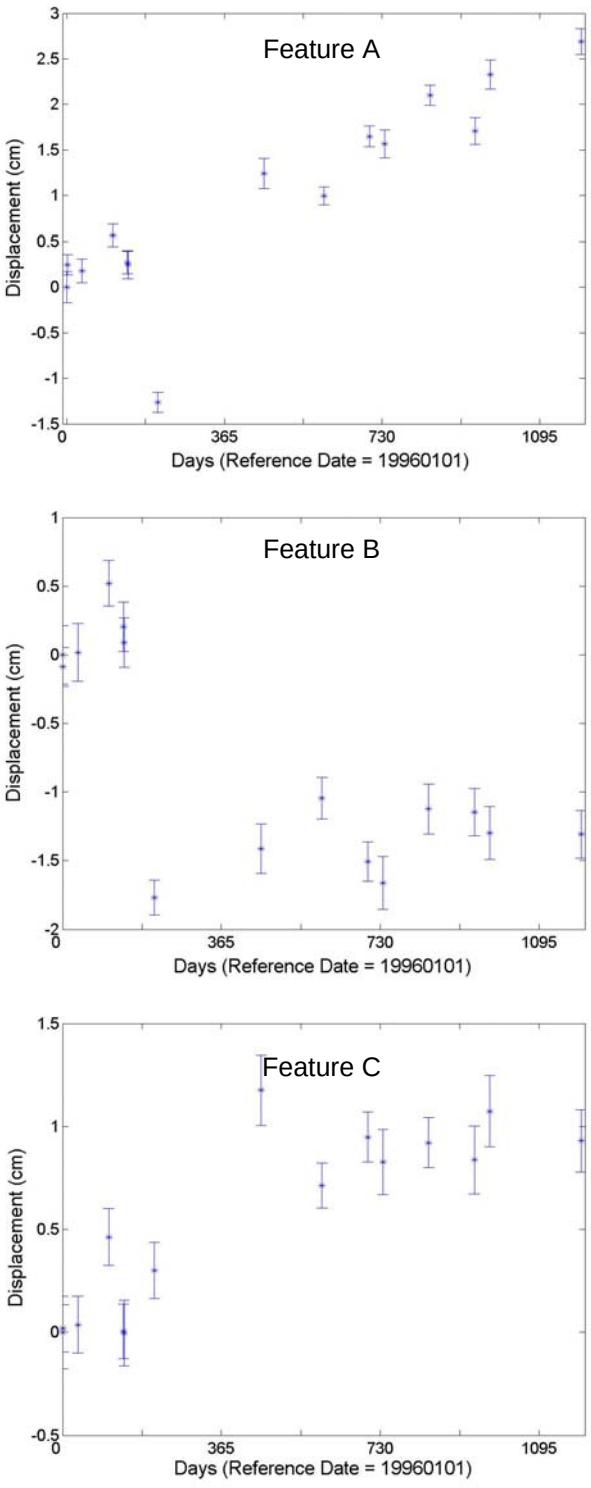


Figure 2. Cumulative displacement along radar line-of-sight since January 1, 1996 for features identified in previous figure. Positive (negative) displacement represents movement toward (away from) radar. Error bars correspond to one standard deviation for interferogram phase noise and SRTM DEM error variances propagated through inversion.

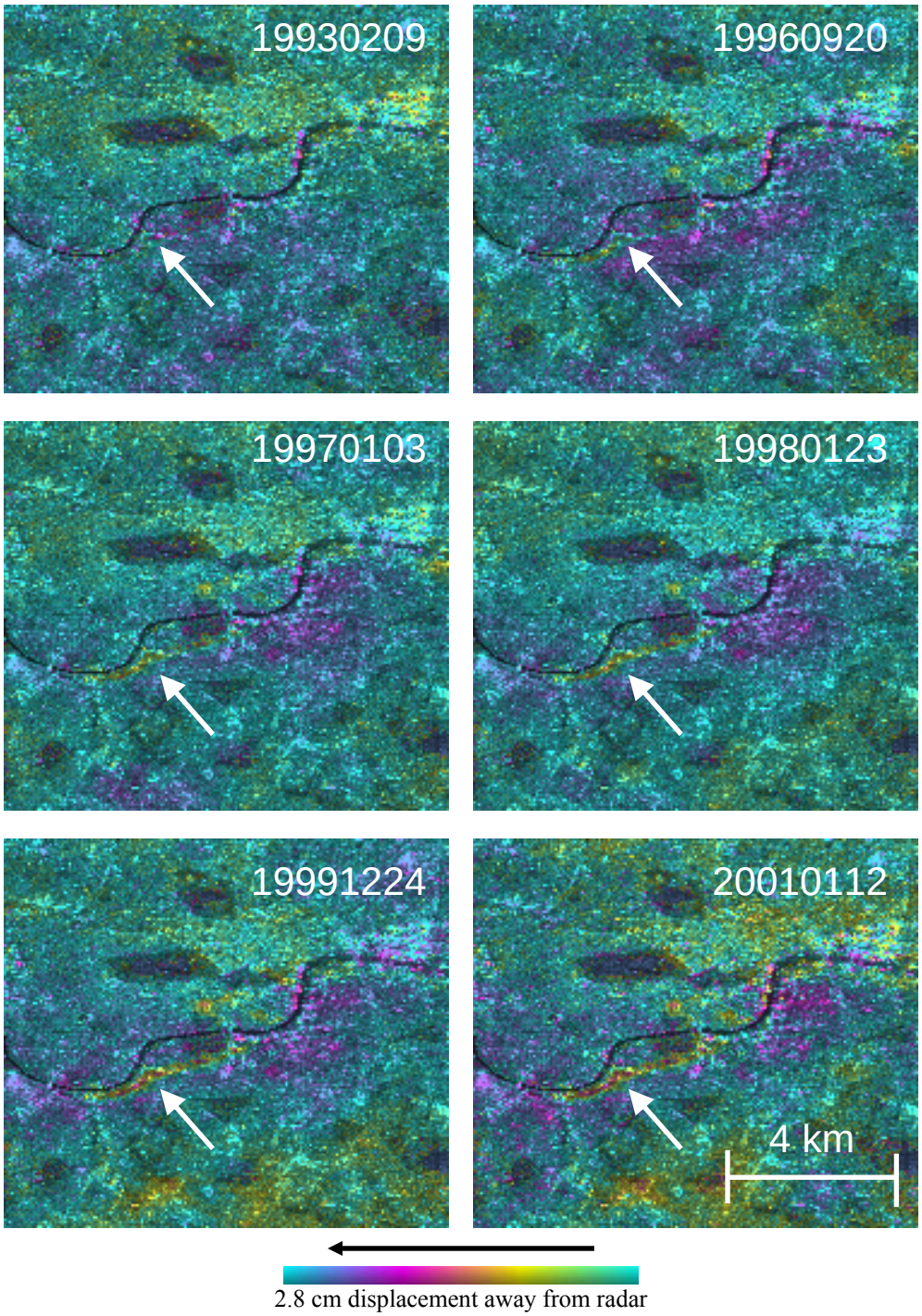


Figure 3. Maps of cumulative displacement since May 5, 1992 for London tunneling area. One cycle through color wheel represents 2.8 cm of radar line-of-sight displacement.

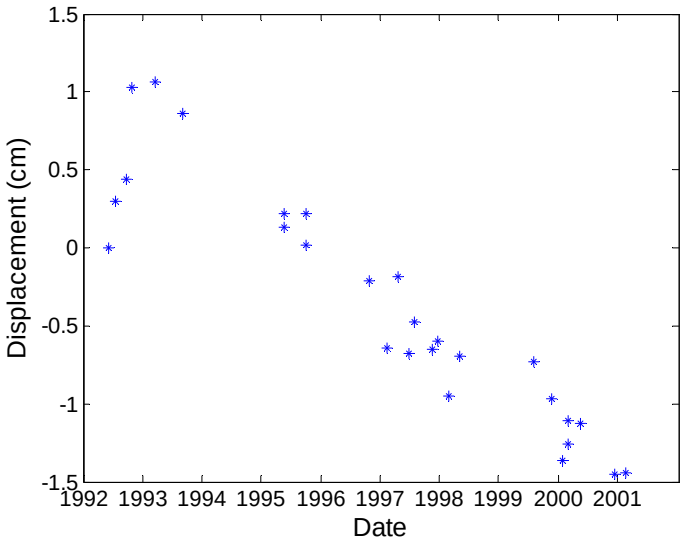


Figure 4. Cumulative displacement along radar line-of-sight since May 5, 1992 for London tunneling area identified by arrow in previous figure. Positive (negative) displacement represents movement toward (away from) radar.

THE 2005 MATSEIS AND NNSA SEISMIC REGIONAL ANALYSIS TOOLS

Darren M. Hart, B. John Merchant, J. Mark Harris, and Christopher J. Young

Sandia National Laboratories

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. DE-AC04-94AL85000

ABSTRACT

In continuing to support the National Nuclear Security Administration (NNSA) Knowledge Base (KB), we developed a set of prototype seismic regional analysis tools (MatSeis, EventID Tool, CodaMag Tool, PhaseMatch Tool, and Infra Tool) to facilitate evaluation of new KB data products, and we continue to maintain and improve these tools. A brief description of the latest versions of MatSeis and the regional analysis tools are given below. MatSeis has been extended to use the FISSURES data handling interface (DHI) client FISSURES-Matlab-Interface (FMI) to connect to both the Northern and Southern California data centers located at Berkeley and Caltech, respectively. A routine was built to support instrument response file download and to automatically perform waveform conversion from counts to nanometers. In addition, both SENSOR and INSTRUMENT flatfile database tables are now exportable through the "waveform write" function. For PhaseMatch Tool, a new method has been added to select station-event pairs for dispersion curve calculations. The user is presented with a list of available stations for a given event, and by selecting from the list, a great circle path is highlighted and the dispersion curve presented. Infra Tool now has the ability to remove a specified signal (i.e., signal-subtraction) from its analysis, in an attempt to detect secondary signals. Preliminary results demonstrate complete removal of the specified signals, allowing for further analysis of residual FK plane. Work is continuing in this area to refine methodology and increase user functionality. For CodaMag Tool, the station-specific MDAC source spectrum and S-wave corner frequency are computed and displayed for reference on the moment spectrum determined by modeling the coda magnitude. Also, Mw is read from any available sources (i.e., STAMAG, NETMAG and DISCRIM_DATA) and displayed as an internal consistency check. For EventID Tool, the amplitude measurement type was switched to the time-domain RMS (TRMS) method, and we have also implemented linear discrimination analysis (LDA), and developed a new tool for computing integrated attenuation maps for a user specified location. In this paper, each of the aforementioned additions are described in greater detail with regard to usage and where relevant, their underlining theory.

OBJECTIVE

MatSeis and the regional analysis tools continue to provide an excellent prototyping environment in which promising seismic analysis techniques can be implemented and evaluated quickly and with relatively little effort. These products are now all fairly mature, but development on them continues in response to the research results from our co-workers at LANL and LLNL and to feedback from our users at other institutions.

RESEARCH ACCOMPLISHED

The following sections describe the latest versions of MatSeis, and the regional seismic analysis tools – PhaseMatch Tool, Infra Tool, CodaMag Tool and EventID Tool.

MatSeis

The main MatSeis graphical window is a standard time vs. epicentral distance plot that can display waveforms, arrivals, origins, and travel time curves. The user can interact with this display by clicking directly on the displayed objects, by using the buttons along the bottom, by using the menus along the top, or by typing commands at the Matlab prompt. MatSeis is predominantly written as Matlab m-file functions, which are organized in a set of directories according to the general purpose of each. However, the package also includes a set of compiled C functions and Java objects. Typically the compiled code is introduced where performance of an m-file is too slow (e.g., FK calculations) or cumbersome (e.g., managing the waveform, arrival, origin, and travel time objects). For a more detailed description of the features in MatSeis, see Hart and Young (2004).

MatSeis continues to be available to the public from the NNSA NEM R&E website (https://www.nemre.nnsa.doe.gov/cgi-bin /prod /nemre/matseis.cgi). We have introduced a new method for tracking downloads of MatSeis that allows us to compile a list of registration information provided by users and create maps of user locations. Our first map of this type is shown in Figure 1, detailing the broad distribution of MatSeis, especially in Western Europe and the east coast of the United States.

Data Center Access

For the last several versions, MatSeis has been reading data from two standard sources: Java database connectivity (JDBC) compliant database tables (e.g., Oracle) and flatfile database tables, each following the NNSA Core Schema (Carr, 2002). In a recent collaboration with University of Washington and IRIS DMC, MatSeis is now able to read data directly from suitably equipped data centers, across the internet. This is possible through FISSURES DHI client services and the DHI client FISSURES-MATLAB-INTERFACE (FMI). Initially, work was done to connect to the IRIS DMC event, network and seismogram server. Once these connections were established and tested, work proceeded to connect to other data centers that were running the same FISSURES DHI client services. These data centers are the Northern California Earthquake Data Center (NCEDC) located at the University of California – Berkeley and the Southern California Earthquake Data Center (SCEDC) located at the California Institute of Technology. Our main goal in this task was to establish connections to the seismogram and network servers for the retrieval of waveform and instrument response data. Over time, the connections were established and routines were developed to use the FMI client as the backbone for download. Once this basic capability was available in MatSeis, we were then able to work on more detailed use cases. One of these was the conversion of retrieved waveforms from counts to nano-meters (nm). In this use case, the user sets up a time-window based query, start time and duration, and specifies the network, station, channel and location code desired for download. The provided information is used by the data center to search its holdings and reply with the number of waveforms that met the provided criteria. The user can use the received information about available holdings to further restrict which waveforms to download, by specifying limits in latitude/longitude coordinates or by distance/azimuth. In the end the user will select to retrieve the waveforms. Once the waveforms are retrieved, the user is prompted to decide whether waveform calibration is desired. If it is, then a list is built of the waveforms network(s), station(s), channel(s) and

location codes and used in querying the data center's network server for the instrument response files. The available instrument response files are then downloaded, saved to disk and then used to calibrate the waveforms to a specific frequency based on the waveform's associated channel type (e.g., UHZ calibration frequency equals 0.004 Hz, BHZ calibration frequency equals 1.0 Hz and HHZ calibration frequency equals 10.0 Hz). The downloaded instrument response files from a data center's network server are in SEED format and are based on the number of stage filters used by the instrument. The final step in this process was to build up routines for generating flatfile tables for SENSOR and INSTRUMENT so the instrument response associations would be available for future use. With these routines in place, a nearly complete set of flatfile database tables (e.g., arrival, assoc, instrument, lastid, netmag, origerr, origin, sensor, site, sitechan, and wfdisc) and data directories (i.e., /resp and /w) can be generated, upon export, after event processing during a MatSeis session.

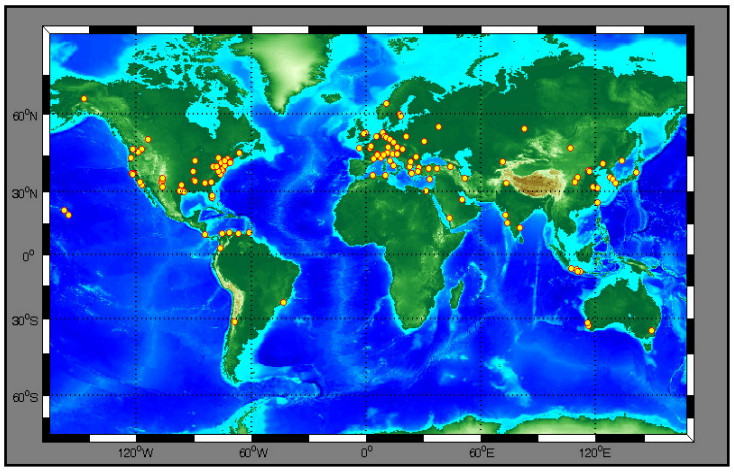


Figure 1. Map of locations for 167 separate MatSeis downloads from December 2004 through July 2005.

Phase Match Tool

PhaseMatch Tool is a waveform analysis interface launched from MatSeis that allows the user to calculate the predicted surface wave dispersion for a given source to receiver path by ray tracing through a model, and then use the model dispersion to generate and apply a matched filter (Herrin and Goforth, 1977). The tool allows the user to view the observed waveform, the model dispersion, the predicted waveform, the cross-correlation of the predicted and observed waveforms, and the match-filtered waveform. The user can control the frequency range of the model dispersion used, as well as the time limit of the portion of the cross-correlated waveform from which the match-filtered waveform is taken. Once a satisfactory filtering has been achieved, the user can send either the observed waveform or the filtered waveform to the MatSeis amplitude measurement widget -- Measure Tool -- to measure surface wave amplitudes. These amplitudes can then be used to determine event magnitude, which in turn can be used as part of an mb/Ms discriminant.

The new surface wave analysis features are part of the LR Path Tool, which we designed as a companion tool to Phase Match to allow the user to examine the dispersion models they are using. By loading an event and set of stations into MatSeis, launching the LR Path Tool and selecting the new button "Set to Map GC", MatSeis takes the selected origin and station list and builds the great circle paths for event-station pairs and presents them in MapTool. Then generates a listbox with station information (i.e., great circle object

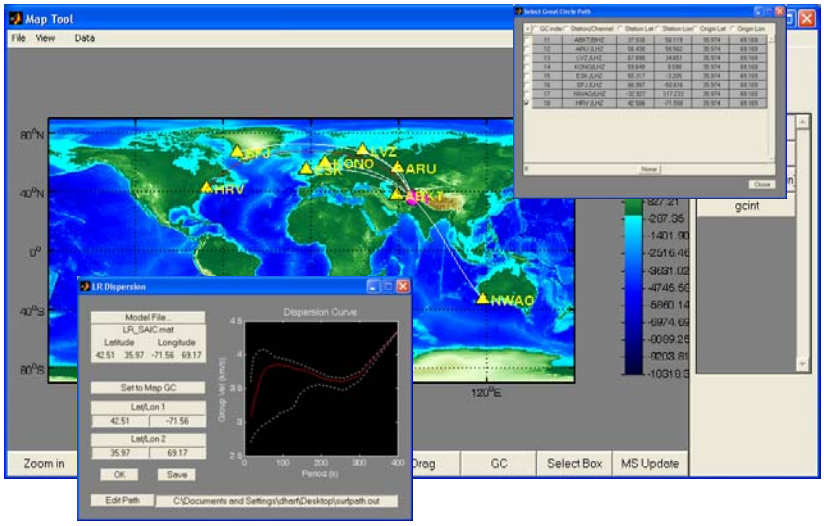


Figure 2. New LR Path Tool. The list of loaded stations is presented and the LR dispersion curve is updated based on station to event path.

index, station channel name, great circle path endpoint coordinates). After selecting a station from the list, the tool computes and displays its dispersion curve. Figure 2 illustrates the set of windows used while working with the new LR Path Tool.

Coda Magnitude Tool

CodaMag Tool is a waveform analysis interface launched from MatSeis that allows the user to calculate magnitudes and source spectra for an event of interest by fitting empirical decay functions to narrow-band coda envelopes of a given phase (currently Lg). The technique was developed by Mayeda and has been described in detail in several papers (Mayeda, 1993; Mayeda and Walter, 1996; Mayeda et al., 1999; Mayeda et al., 2003). The tool consists of two displays. The main one shows the calculated moment spectrum and the derived magnitudes. The second display, launched from the first, shows how the spectrum was derived. The user can adjust the Lg arrival window, examine the fit between the observed and synthetic envelopes, and control which frequency bands are used for the magnitude calculations. The various required parameters (frequency bands, groups velocity windows, decay curves, etc.) are read from parameter files unique to each station.

New features to version 1.10 include display of theoretical source moment–rate spectrum and S-wave corner frequency to help evaluate the quantities derived from the data. For computing the CodaMag theoretical source moment-rate spectrum, we started with the spectral shape defined by Brune (1970) given in Equation 1.

$$S(w) = \frac{So}{1 + (\frac{w}{w_c})^2} \qquad \text{Equation 1.}$$

Where S_0 is the zero frequency spectral level and w_c is the angular corner frequency. Since the moment-rate spectrum is required to have a zero frequency spectral level equal to the moment (M_0), the term S_0 in equation 1 is substituted with the moment, yielding Equation 2, which is the moment-rate spectrum plotted in CodaMag Tool.

$$\text{moment-rate spectrum}(w) = \frac{Mo}{1 + (\frac{w}{w_c})^2} \qquad \text{Equation 2.}$$

For calculation of the S-wave corner frequency we use the formulation by Walter and Taylor (2002), using station specific Magnitude-Distance Amplitude Correction (MDAC) parameters. Figure 3 shows a screen-shot of version 1.10 of CodaMag Tool, illustrating the new features of the tool.

The upcoming work on CodaMag Tool will include displaying values of Mw retrieved from different internal database sources (i.e. STAMAG, NETMAG and DISCRIM_DATA tables) to help evaluate the accuracy of coda magnitude calibration for a given station, multiple-band yield estimates by region, reading processing parameters from database tables, and storage of the processing parameters in database tables.

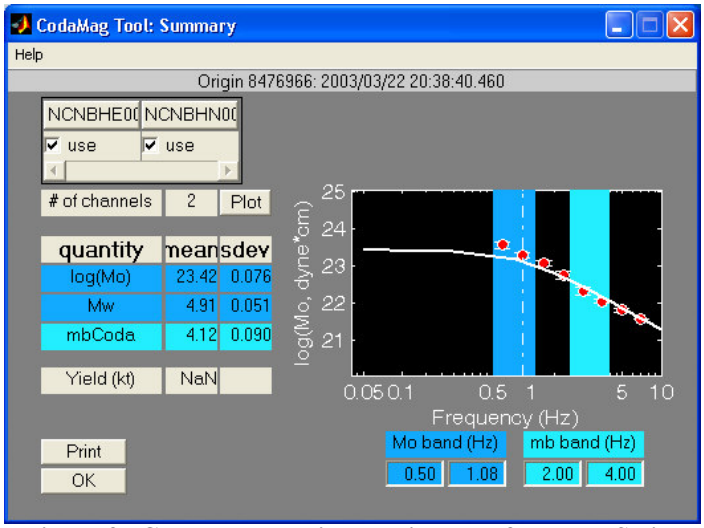


Figure 3. CodaMag Tool illustrating new features. Solid white line is the moment-rate spectrum, and vertical white dashed line is the S-wave corner frequency (f_{cS} = 0.84 Hz). Processing parameters provided by Kevin Mayeda of Lawrence Livermore National Laboratory.

Event Identification Tool

EventID Tool is a waveform analysis interface launched from MatSeis that allows the user to categorize an event of interest (i.e. explosion or earthquake) using spectral ratios of standard regional arrivals (see Hartse et al., 1997; Walter et al., 1999). This is a station-calibrated technique, and researchers must provide a set of station-specific calibration parameters that are stored in a set of database tables and assessed by the tool when processing an event of interest. The tool consists of three displays. The main display plots the phase ratio for the current event against a backdrop of the same ratio for archived events (if available) that have already been identified (event types are indicated with different symbology). The user can choose different phases and/or frequency bands to ratio to try to improve the separation of the earthquake and explosion populations, and the display will immediately update. A second display shows the user a plot of an "MDAC-o-gram" (i.e. the MDAC corrected measurements at all of the phase/frequency combinations) for the current event along with all of the archived events. This can be useful in deciding which ratio will yield the best separation. If there are questions about the amplitude measurements themselves, a third display can be brought up, and the user can easily examine group velocity windows for the phases and change them if necessary. If they are changed, the measurements will automatically be re-made and the ratios will be updated in the main display.

Version 1.10 of EventID Tool supports both time and frequency domain amplitude measurement types, as defined in the custom schema for DISCRIM_DATA (i.e. FREQ and TRMS). It also implements the linear discrimination analysis (LDA) technique, can run validation scripts, and can generate path-integrated attenuation (Q) maps. We describe these new features below.

In previous versions of EventID Tool, amplitude measurements were made in the frequency domain (i.e. FREQ f_t_type measurements), while the reference events were being read from the database based on their time-domain (i.e. TRMS f_t_type measurements) counterparts. In theory, the measurements are equivalent, but in practice, slight differences are observed, so a consistent method of amplitude measurement was desired. The tool will now compute both measurement types for the current event, and comparisons can be made using the MDAC-o-gram, as shown in Figure 4. Future work will focus on allowing more flexibility within the tool for switching between measurement types, both for validation and operational purposes. For a more detailed description of the TRMS and FREQ measurement techniques see Rogers et al. (2002).

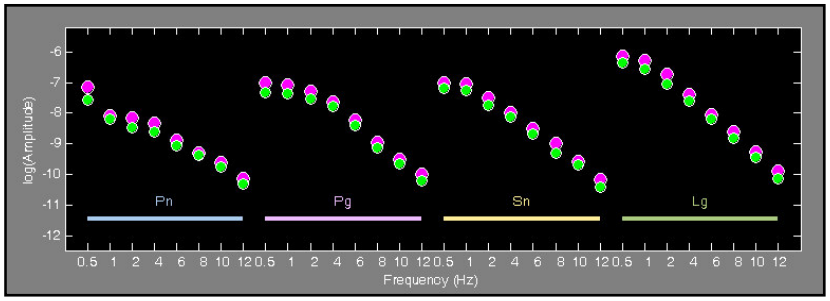


Figure 4. Plot demonstrating log(amplitude) measurement types FREQ (green) and TRMS (magenta) across 8 frequency bands (0.5-1 Hz, 1-2 Hz, 2-4 Hz, 4-6 Hz, 8-10 Hz, 10-12 Hz and 12-16 Hz). In the case shown here, the log residual differences (i.e. TRMS-FREQ) had a standard deviation of 0.113 about the mean of 0.211 for Pn, 0.066 about the mean of 0.207 for Pg, 0.057 about the mean of 0.194 for Sn and 0.067 about the mean of 0.207 for Lg.

For evaluation of the LDA technique, new controls were added to implement this method as unobtrusively as possible without interfering with the existing single-ratio discriminant capability. As seen in Figure 5, the previous methods for selecting phase, spectral and cross-spectral ratios are located just above the new section for selecting and editing a LDA string. The user controls the active method by selecting from a radio-button, "on" or "off", switch. Upon selecting one of the discrimination methods the other methods controls are disabled ensuring that the tool and user are clear as to the desired functionality. There are many ways to implement LDA within the tool; we anticipate further work on refining this interface based on feedback from users.

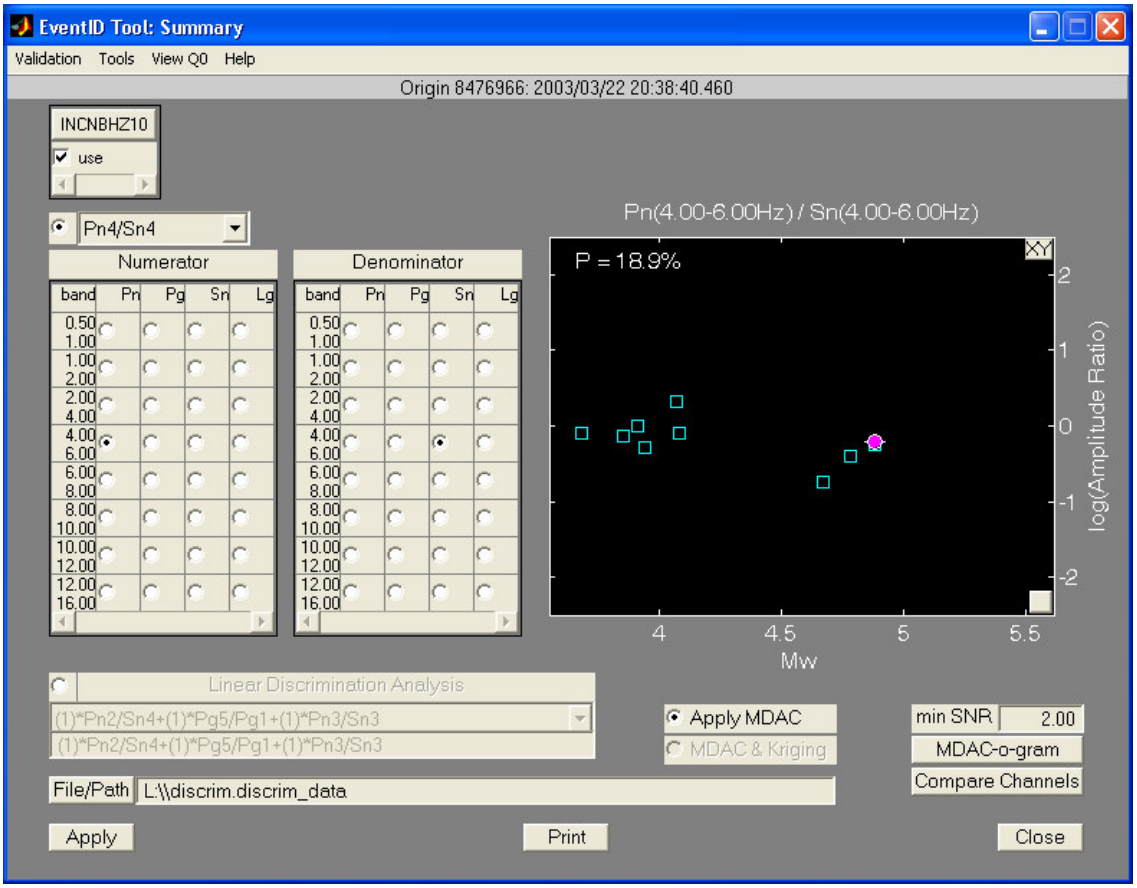


Figure 5. EventID Tool. Minor layout changes have been made to allow for new LDA controls. Menus have been added for selection of other new feature, e.g., validation methods, Q model viewer tool. Processing parameters were provided by Bill Walter of Lawrence Livermore National Laboratory. A hypothetical LDA string is shown to demonstrate the format intended for use by the tool.

To help researchers quickly validate ever larger numbers of calibrated stations, we have introduced a new capability into the tool. A routinely used script for product validation was modified to handle an abundant number of cases that the data products might fall within. The output from the validation script provides useful numerical analyses for an event that has been delivered in the DISCRIM_DATA database table and was processed by the tool using data from DISCRIM_PARM, MDAC_FI and MDAF_FD. Comparisons are separately made between the reference event and tool processed event, for both the MDAC corrected and uncorrected log(amplitude) measurements, providing residuals for evaluation of measurement consistency. Other useful information is also output; e.g., station and channel name, event type, start and end time for phase processing windows. Additional possible output variables may be added if requested.

To better understand the effects of attenuation observed at a station for events occurring at different source points, a new tool was developed to quickly build path-integrated Q maps using the 2D Q models that are included with some calibrated stations. A path-integrated Q map consists of taking an existing tomography grid of Q values, defining a point of interest within the Q tomography grid (in this case the station of interest), defining a path-integration grid for the integration to take place over, and then computes the path-integrated Q values for the great-circle paths between the station and each of the path-integration grid points. An example path-integrated Q map is shown in Figure 6, for station ARA0 of the ARCESS array, using a 1Hz Pn Q tomography model.

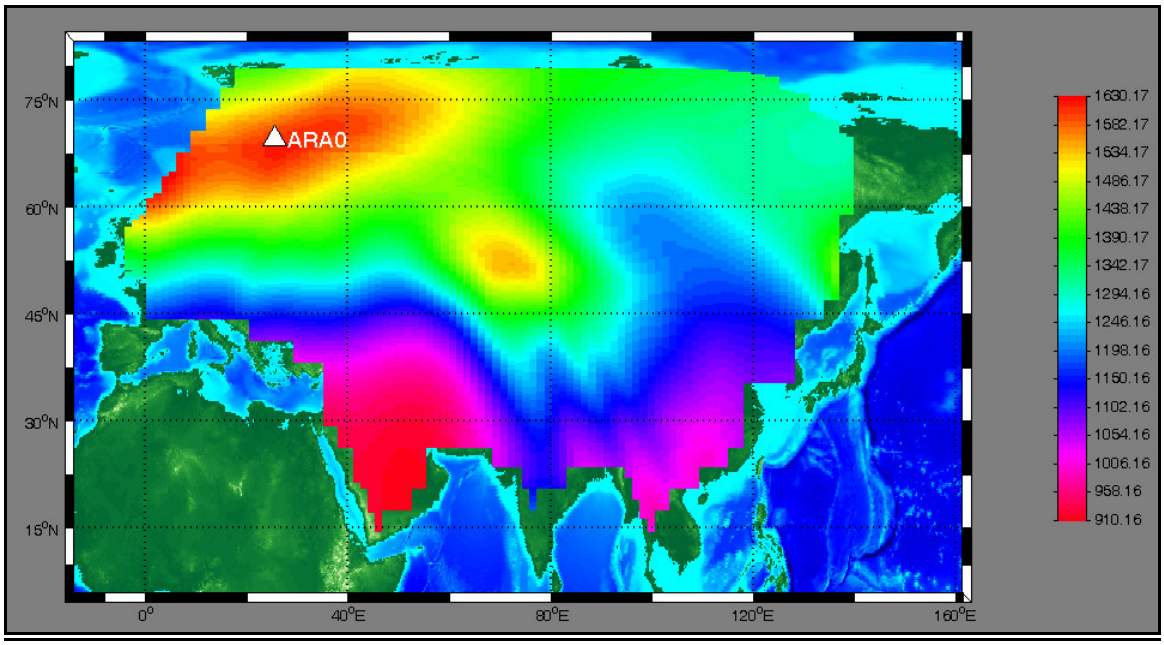


Figure 6. Pn path-integrated Q0 map station ARA0. A grid increment of 0.66 degree in both latitude and longitude was used. Q0 map used to construct this plot was provided by Steve Taylor of Los Alamos National Laboratory.

Infra Tool

Research and evaluation is currently underway to extend Infra Tool for 1) detection of imbedded secondary signals, and 2) removal of a constant noise source from infrasound array data. For the first task a signal-subtraction technique is used for removal of the primary source of correlated signal power from the FK plane for each window step when processing a segment of data. Our implementation of signal-subtraction is done by beamforming the infrasound array to the azimuth and slowness defined by the peak correlated power in the FK plane, then subtracting the beam from each array element and re-computing the FK on the beam-subtracted waveforms. After signal subtraction, the residual FK plane will have a new point of peak correlated power from which detection criteria (Hart, 2004) can be used to determine if secondary signals are present. For the second task of constant noise source jamming, we use a fixed azimuth and slowness for beamforming and the residual FK plane is again searched for signals meeting the detection criteria. Our purpose is to provide functionality to remove a constant noise source (e.g., microbaroms), or any peak correlated power observed in the FK plane (e.g., chemical explosion or artillery blasts). A new GUI, shown in Figure 7, was designed and built to control the signal-subtraction processing. Features in the new GUI include methods to switch between fixed source jamming and peak correlation, a switch to turn ‘on’ and ‘off’ results from non-signal-subtraction, and a switch for applying or not applying the signal subtraction method. Work continues on developing evaluation plots of the technique for the user to review for validity of the method.

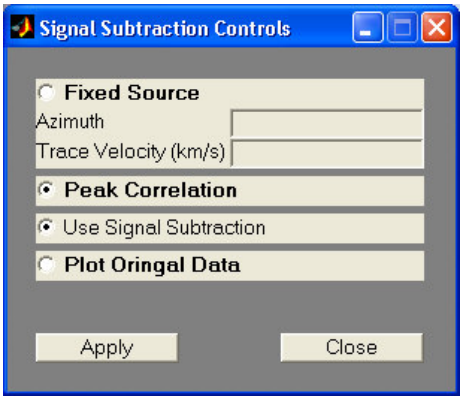
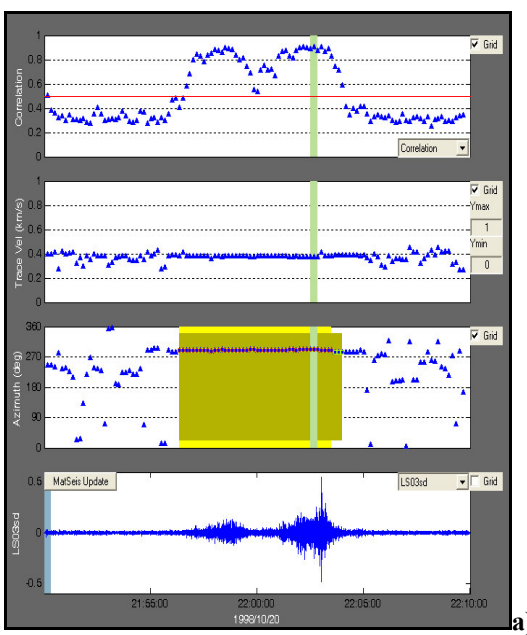
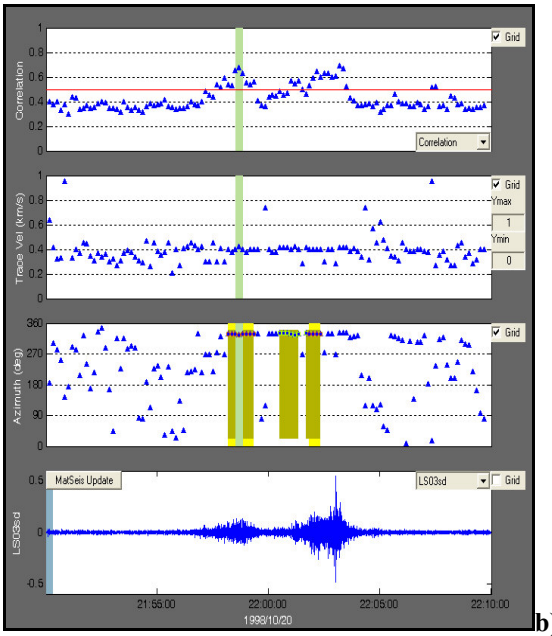


Figure 7. New graphical user interface for the signal-subtraction method planned for Infra Tool. Select fixed source and add specific azimuth and trace velocity values for the fixed noise source jamming method, or choose the peak correlation method. Select the Use Signal Subtraction to turn "on" and "off" the signal-subtraction processing routine. By selecting the Plot Original Data, the results from non-signal subtraction are displayed for comparison.

As an example, we show results from processing using the standard method and signal-subtraction for peak correlated power removal and secondary signal detection in Figures 8 a and b. Figure 8a shows that the standard processing method clearly detects two pulses with start times of 1998/10/20 21:56:20 and 1998/10/20 22:01:00 and durations 3.3 minutes and 2.5 minutes, respectively. The average azimuth of these pulses is 289.2 degrees. Figure 8b shows the signal-subtraction results of the initial peak correlated power obtained by the standard processing method. In this case three detections are made and each occurs at the approximately same azimuth, 327.2 degrees. These are likely refracted wavefronts from local topography near the infrasound station of the primary pulses detected with the standard processing method.



a)



b)

Figures 8a and 8b. Demonstration of signal-subtraction technique for removal of peak correlated power and secondary signal detection. A waveform segment with start time 1998/10/20 21:50:00 and 20 minute duration was chosen for the Los Alamos Infrasound Array (LSAR). Waveforms were bandpass filtered at 1-3 Hz (bottom displays of both panels) and processed using 20 second moving window with 50% overlap. Detection threshold criteria were set to: correlation > 0.5, slope limit in azimuth < 1.0 degrees, Number of samples 4 (i.e. 80 second LTA), standard deviation in azimuth < 2.5 and azimuth data gap 0. a) Standard processing results, with associated detection windows. Detection window statistics: average correlation 0.8 ± 0.097, trace velocity 380 m/s ± 7 and azimuth 289.2 ± 2.0. b) Signal-subtraction results for removal of peak correlated power, with associated detection windows. Detection window statistics: average correlation 0.566 ± 0.105, trace velocity 399 ± 20 m/s and azimuth 327.2 ± 1.8. Data obtained from the U.S. Army Space and Missile Defense Command (SMDC) Monitoring Research website Infrasound database of ground truth events.

CONCLUSIONS AND RECOMMENDATIONS

In this paper, we have highlighted some of the more significant changes in the past year in MatSeis and the MatSeis-based regional analysis tools. In the coming year and beyond, we expect to continue to modify the tools in response to user feedback, with the ultimate goal of building a well-integrated, user-friendly, and powerful regional analysis environment.

The official 1.9 version of the basic MatSeis package is available to all from the GNEM R&E web site: https://www.nemre.nnsa.doe.gov/cgi-bin/prod/nemre/matseis.cgi, with MatSeis version 1.10 planned for release in November 2005.

Matlab and the Signal Processing Toolbox are required to run MatSeis. MatSeis will run on Sun workstations, Windows PC's, and Linux PC's. MatSeis should run on other supported Matlab platforms as well, but the C code will need to be re-compiled.

ACKNOWLEDGEMENTS

We thank all of the MatSeis users who have helped us to debug and improve the software, particularly our colleagues at LANL and LLNL.

REFERENCES

Carr, D. (2002), National Nuclear Security Administration Knowledge Base Core Table Schema Document, Sandia National Laboratories, SAND2002-3005 (available at http://www.nemre.nn.doe.gov).

Hart, D. M. (2004), Automated Infrasound Signal Detection Algorithms Implemented in MatSeis – Infra Tool, Sandia Technical Report SAND2004-1889.

Hart, D. M. and Chris J. Young (2004), MatSeis User Manual version 1.9, https://www.nemre.nnsa.doe.gov/prod/nemre/fileshare/matseis_manual.pdf.

Hartse, H. E., Taylor, S. R., Phillips, W. S., and G. E. Randall (1997). A preliminary study of regional seismic discrimination in central Asia with emphasis on western China, *Bull. Seism. Soc. Am*. 87: 551-568.

Herrin, E. and T. Goforth (1977). Phase-matched filters: application to the study of rayleigh waves, *Bull. Seism. Soc. Am.* 67: 1259-1275.

Mayeda, K. (1993). Mb(LgCoda): a stable single station estimator of magnitude, *Bull. Seism. Soc. Am*. 83: 851-861.

Mayeda, K. and W. R. Walter (1996). Moment, energy, stress drop, and source spectra of western United States earthquakes from regional coda envelopes, *J. Geophys. Res*. 101: 11,195-11,208.

Mayeda, K., R. Hofstetter, A. J. Rodgers and W. R. Walter (1999). Applying coda envelope measurements to local and regional waveforms for stable estimates of magnitude, source spectra, and energy, *21st Annual Seismic Research Symposium on Monitoring a CTBT*, Vol I, pp. 527-533.

Mayeda, K., A. Hofstetter, J. L. O'Boyle and W. R. Walter (2003). Stable and Transportable Regional Magnitudes Based on Coda-Derived Moment-Rate Spectra, *Bull. Seism. Soc. Am.* 93: 224-293.

Rogers, A. J., T. Lay, W. R. Walter, and K. M. Mayeda (1997). Comparison of Regional Phase Amplitude Ratio Measurement Techniques, *Bull. Seism. Soc. Am.*, 87, 1613-1621.

Walter, W. R., A. J. Rodgers, A. Sicherman, W. Hanley, K. Mayeda, S. C. Myers and M. Pasyanos (1999). LLNL's regional seismic monitoring research, *21st Annual Seismic Research Symposium on Monitoring a CTBT*, Vol I, pp. 294-302.

Walter, W.R. and S. R. Taylor (2002), A Revised Magnitude and Distance Amplitude Correction (MDAC2) Procedure for Regional Seismic Discriminants: Theory and Testing at NTS, Lawrence Livermore National Laboratory, Livermore CA, report: UCRL-ID146882 and Los Alamos report: LA-UR-02-1008.

DEVELOPMENT OF A NETWORK DATA SET FOR EVALUATING DETECTION AND NETWORK PROCESSING PERFORMANCE

Benjamin Kohl, Theron J. Bennett, István Bondár, Brian Barker, Walter Nagy, Colin Reasoner, Hans Israelsson, and Paul Piraino

Science Applications International Corporation

Sponsored by Army Space and Missile Defense Command

Contract No. DASG60-03-C-0009

ABSTRACT

Practical implementations of new or improved monitoring technologies, such as signal detectors, network phase association algorithms, location and event identification methods, rely on quantitative assessments of performance such as detection probabilities and false alarm rates. These types of performance metrics are typically obtained through experiments using data sets constructed from archival data records. However, such experimental data sets implicitly contain signal and event recordings from numerous unknown sources (e.g., small earthquakes not reported in published local, regional, or teleseismic bulletins) potentially contaminating the data set and complicating the interpretation of processing results. Furthermore, they are only representative of events and station network characteristics contained during the time interval of the archival data. Our objective is to develop an experimental network data set in which all the target signal and event detections are known and ultimately to extend those results to represent expected network data from potential surrogate events and stations, which may not be included in the historical archive. To achieve this objective, we have been developing the framework for synthesizing a database including continuous waveform data for a network of seismic and infrasound stations relevant to nuclear explosion monitoring which contains signals from actual events, scaled to various sizes, and embedded in a variety of background noise.

Our initial focus for this study has been a large region in southern Asia (15°-45°N 50°-115°E). We have identified a network of 51 core seismic and infrasound stations, most useful for monitoring this region; and we have been collecting waveform data from those stations to represent background noise and signals from historical nuclear explosions as well as earthquakes and seismo-acoustic sources. In constructing the data for background noise, we are seeking to form long, continuous waveforms of detection-free clean noise spanning several days into which we can then embed real event signals and signals which have been scaled down on the basis of source scaling predictions to magnitudes representing lower levels. Formation of clean noise waveforms has required meticulous analysis to exclude time-windows with phase arrivals predicted from global and regional seismic bulletins as well as phases picked by standard signal detectors. Resulting noise segments have been carefully merged together to produce several days of continuous clean noise waveforms while maintaining basic noise attributes with respect to overall level and seasonal, weekly, and diurnal variations. From our effort to date, we have generated clean noise waveforms of two-days duration, as well as reversed noise waveforms of similar duration, for 42 of the seismic stations.

We have assembled the seismic signal waveforms from 6 underground nuclear explosions and approximately 100 well-recorded earthquakes with high signal-to-noise ratio (SNR) which occurred in southern Asia along with seismo-acoustic signals from 23 mine blasts and one bolide recorded by infrasound stations in Mongolia and Kazakhstan. We have been testing and employing frequency-dependent explosion (e.g., Mueller/Murphy) and earthquake (e.g. Brune with both inverse cube- and quad-root corner frequency dependence on moment) source scaling models to scale down the large, high-SNR events to small events covering a range of yields/magnitudes approaching the monitoring thresholds. In addition to describing target events for analyzing monitoring performance, the scaling/embedding process is also being used to represent potential sources of regional and teleseismic clutter signals, which increases processing complexity (while continuing to maintain control of the contributing sources) and provides a more realistic background condition than the clean noise scenario. Preliminary event detection experiments are quantifying the systematic time, amplitude and azimuth measurement biases that can be expected from low-SNR detections. Methodologies for analyzing the performance of the infrasound stations for monitoring seismo-acoustic events from the southern Asia source region are also being assessed.

OBJECTIVES

Our objective is to develop an experimental network data set in which target signal and event detections are known, as well as having realistic distributions of false or "clutter" detections and background noise characteristics. We are utilizing a variety of actual nuclear explosion, earthquake, mine blast and infrasonic event recordings and developing scaling and embedding algorithms to yield continuous waveforms with numerous target events at or near the detection threshold in southern Asia. This will allow for detection, location and identification experiments utilizing the known characteristics of small events under realistic background noise and seismicity conditions. The background noise, scaled signals, embedded waveforms and relevant meta-data from this effort are available to the monitoring research community via the mechanisms of the Research and Development Support Services (RDSS) web site (http://www.rdss.info/, Woodward et al. 2005).

RESEARCH ACCOMPLISHED

The basic framework for the comprehensive network data set was developed previously (Kohl et al., 2004) and is depicted in Figure 1. It involves taking well-recorded, high signal-to-noise ratio (SNR) signals, scaling them down to various sizes based on source theory and embedding them in a variety of background noise conditions. We are scaling and embedding nuclear explosion, earthquake, seismo-acoustic (e.g., mine blast) and infrasonic event recordings at levels spanning the detection threshold.

We assembled a background noise library of two days of continuous detection-free clean noise, two days of reversed clean noise and fours days of whole background noise for a core network of 42 stations. We constructed a signal library from 6 nuclear explosions, over 100 earthquakes, 23 mine blasts and one bolide. We developed source scaling models for nuclear explosions and earthquakes and scaled the nuclear explosion records and earthquakes to equivalent m_b ranging from 1.8 to 4.5. We embedded the scaled signals several hundred times in varying noise conditions and conducted a number of signal detection experiments yielding Receiver Operating Characteristic (ROC) curves and quantitative assessments of arrival time, azimuth and amplitude biases as a function of SNR for selected stations. We continue to develop scaling models for infrasound signals and plan on conducting experiments to demonstrate the utility of this approach to assess network processing performance.

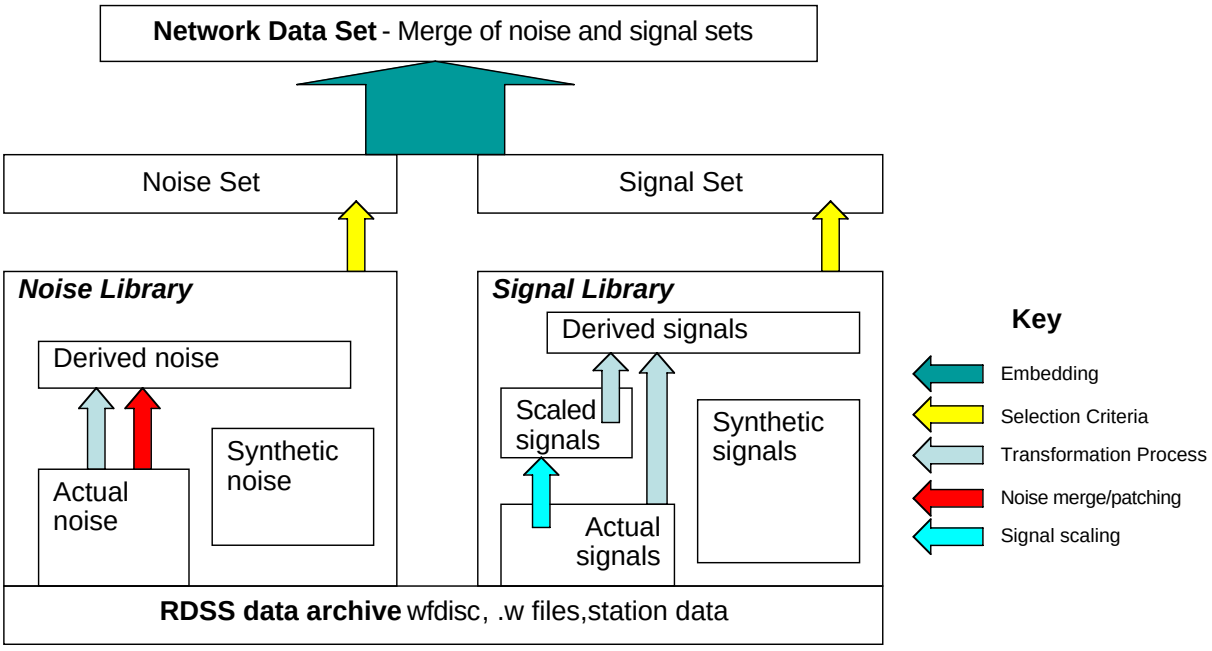


Figure 1. Basic framework for constructing a network data set for systematic testing and evaluation of new or improved monitoring technologies. Scaled signals from a variety of sources are embedded in a variety of noise conditions to construct a data set in which target detections have been characterized.

Noise and Signal Library

In the past year we completed the construction of three background noise datasets for a network of 42 stations useful for building experimental data sets for testing monitoring capabilities in southern Asia (15°-45°N 50°-115°E). The first of these data sets (clean-noise, two days) was constructed by stitching together detection free segments. To insure that the stitched clean-noise realistically represented actual noise, including known seasonal and diurnal variability, we first assembled a set of reference spectra for each station and channel spanning the range of variability for each station (Figure 2). For example, for stations that exhibited strong diurnal variability in the noise levels, we computed a separate reference spectrum for every hour of the day. Only those detection-free segments derived from the same hourly span, *and* whose spectra matched the reference spectra were used. To minimize the effects of the merging process we used 10 to 120 second tapers at the ends and overlapped neighboring segments. All the channels of stations and arrays were merged consistently in time to retain the noise coherency characteristics originally present in the data.

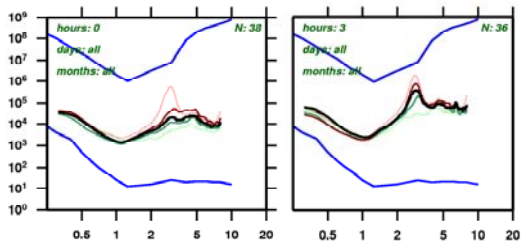


Figure 2. Example reference spectra for hours 0 and 3 for the CM16 element of CMAR.

The second noise dataset (reversed, two days) was constructed by simply time-reversing the clean background noise, thus also yielding two days of continuous noise. The third noise dataset (whole-background) was constructed by simply extracting four days of raw continuous data (June 1998). This third dataset is analogous to what is normally used in signal processing experiments. As a quality control measure, and to establish a baseline background detection rate, we ran standard signal processing (DFX) against the clean noise. Despite the fact that the clean noise was constructed from detection-free segments, low-level detections were still made against the clean and reversed noise. On average the detection rate against the clean-noise was 20% of that against whole-background noise and reversed noise had a detection rate of about 30% of the whole-background noise. Figure 3 shows the waveform and the spectrogram of a waveform stitched from detection-free segments.

In the past year we assembled a signal library from the waveforms of 6 historical nuclear explosions, more than 100 earthquakes, 23 mine blasts and one bolide that occurred in central and southern Asia. We are computing a wide variety of signal characteristics (e.g. Figure 4) on the original event records, and the scaled signals.

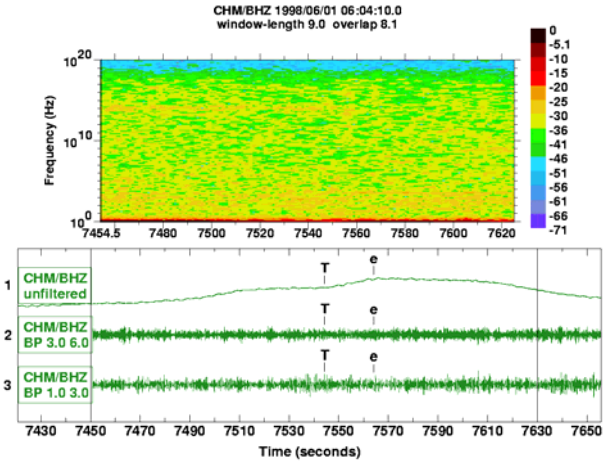


Figure 3. Spectrogram of merged clean-noise around a stitching boundary (marked start – "T" and end – "e" markers).

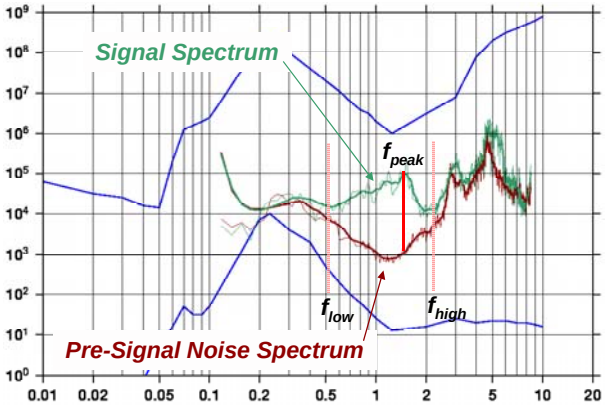


Figure 4. Schematic showing the signal bandwidth and peak-frequency measurements routinely computed for the signal library.

Explosion and Earthquake Scaling

One of the principal objectives of this project is to provide realistic assessments of monitoring performance for smaller events for which the signal detections from station networks are often incomplete. To better understand the factors affecting this performance for our southern Asia study area, we have been using source scaling theory to scale the signals from larger events. This scaling introduces a frequency-dependent change in signal amplitudes.

For explosion scaling we have been using the Mueller-Murphy (MM) model which has been validated over the years for a range of explosion observations (Mueller and Murphy, 1971; Murphy 1977). Alternative models (e.g., vonSeggern and Blandford, 1972) would be expected to produce very similar predicted behavior. The MM model is formulated in terms of explosion yield (W) and provides an expression of the P-wave spectrum as a function of source media properties, explosion yield, depth, and empirical constants corresponding to different geologic emplacement media. For southern Asia we use the explosion scaling relations for granite, which worked well for nuclear explosions at Semipalatinsk and Lop Nor test sites. For scaling the explosions in terms of body-wave magnitude m_b, we use the relation $m_b = 4.45 + 0.75 \log W$, which has been previously validated for these test sites, to convert to yields. Our MM model results were verified by comparing observations of Pn spectral ratios from nearly co-located nuclear explosions with the predictions based on the source scaling theory. Some of these comparisons showed very good matches, although in other cases there appeared to be corner frequency differences, which may require future modifications to some of the model parameters.

To test the explosion scaling model, we applied the MM scaling procedures to scale the signals recorded on a network of regional and teleseismic stations from 6 southern Asia underground nuclear explosions, down from their original magnitudes ($4.5 \le m_b(\text{REB}) \le 6.0$) to a range of lower magnitudes ($m_b(\text{REB}) = 4.5$ and in 0.1 magnitude unit steps from 4.0 m_b(REB) to 1.8 m_b(REB)). We then applied standard IDC processing to the scaled signals to measure initial P amplitudes and associated periods, which would be used for computing station magnitudes. The results are presented in Figure 5a, in which we plot the observed magnitude differences at each stations, $\log A_i/T_i$(original unscaled) – $\log A_i/T_i$(scaled) where i is a station index, versus the target m_b difference, m_b(original for the network) – m_b(target for the network). Obviously the ideal result would be for the observed magnitude difference at each station to equal the target magnitude difference, and this is achieved quite well in the explosion scaling measurements in Figure 5a. The observations are scattered around a line with a slope approximately equal to 1.0. With the exception of a few outliers related to data quality issues, the scatter in the observations is less than half of a magnitude unit, and the least-squares linear fit to the observations is 1.02 with only a slight bias indicated at the largest target magnitude differences. Thus we conclude that the MM explosion scaling procedure appears to be performing as expected over a fairly large range of magnitudes.

For earthquake scaling, we began with a Brune ω^2-source model. For this model the corner frequency is proportional to velocity of the source medium and inversely proportional to a source dimension term, which scales with moment. In scaling the earthquake signals, we consider both cube-root (as indicated in original models by Brune 1970, Hanks and Bakun 2002) and quad-root (as suggested by Mayeda and Walter 1996 amongst others) for

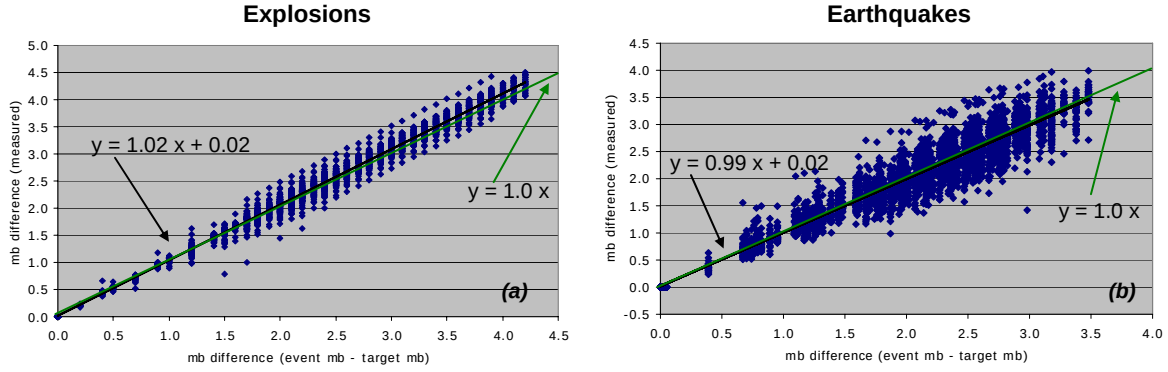


Figure 5. Results of magnitude differences for P and Pn signals measured from routine processing of the observations from 6 southern Asia nuclear explosions scaled using MM explosion source scaling (a) and from 17 southern Asia earthquakes scaled using our preferred cube-root earthquake model (b).

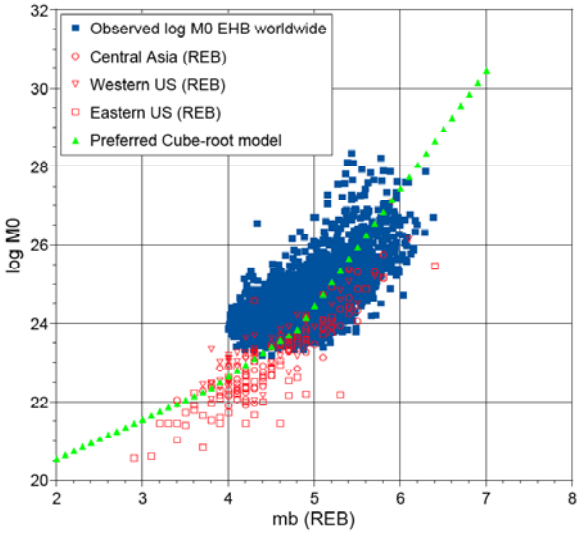


Figure 6. Log M0 versus mb(REB) for a large global earthquake sample reported by EHB and elsewhere.

corner frequency dependence on seismic moment. Our analyses to date focus on the model with cube-root corner frequency dependence. In the earthquake model we followed the approach of Hanks and Bakun (2002), using the definition of moment magnitude, M_w, to establish the relationship to source moment – i.e., $\log M_0 \equiv 1.5\ M_w + 16.05$. We began with a simple linear model for relating M_w to m_b. We drew upon observations from the global EHB (Engdahl et al. 1998) earthquake sample for large events (m_b(REB) > 4) along with observations from smaller events for selected areas reported by Patton (2001), which were adjusted to equivalent REB m_b's (assuming $m_b \approx m_b$(REB) + 0.3). Although the data scatter (Figure 6) would appear to permit a linear M_w-vs-m_b model (i.e. essentially a straight-line relationship between $\log M_0$ and m_b), the implied effects on spectral behavior are not realistic. In particular, for small events with magnitudes measured from signals with frequencies below the corner frequency, a one unit change in m_b should correspond to a factor of ten change in M_0. To meet this objective we would need to have $M_w \propto 2/3 m_b$ – i.e. $\log M_0 \propto 1.0 m_b$. The observations in Figure 6 cannot support such a slope over magnitudes $3 \le m_b(\text{REB}) \le 6$. We, therefore, decided to investigate a model for which the M_w-vs-m_b relationship is nonlinear; Taylor et al. (2002) reached a similar conclusion and developed a model only slightly different from those described below. In particular, we sought to determine an earthquake model for which $\log M_0$ scales directly as the m_b difference for small events, produces $M_w \approx m_b$ over some intermediate range in magnitude, and has M_w greater than m_b for larger events (where we know m_b is saturated).

As a preliminary validation test to further constrain our earthquake model, the earthquake scaling relations were used to scale down the signals from selected samples of large southern Asia earthquakes which were well recorded at regional and teleseismic stations. Just as for the scaled explosion signals above, we compared the A/T measurements from the scaled P and Pn signals to the target values for the earthquakes, m_b(REB) = 2.5 – 4.5. When this scaling model was applied to the signals from 17 southern Asia earthquakes, the A/T measurements were in very good agreement with the expectation (Figure 5b). There appears to be very little indication of bias in the results; the slope of the least-squares straight line fit to the observations is 0.99 and the intercept is 0.02. Scatter about the line is about 0.7 magnitude units, only slightly greater than the scatter seen in the corresponding explosion measurements. We conclude that the cube-root model combined with our $\log M_0 - m_b$ relation (Table 1), provides a reasonable procedure for scaling the southern Asia earthquakes. We are continuing to look at additional validation with spectral ratios (e.g., Figure 7) from nearly co-located events and to evaluate alternatives, including an earthquake model with quad-root corner frequency dependence.

Table 1: Preferred moment-vs-m_b relationship

$\log M_0 - m_b$ relation for cube-root model	
$\log M_0 = m_b$(REB) + 18.55	m_b (REB) ≤ 3.8
$\log M_0 = 1.5\ m_b$ (REB) + 16.65	3.8 ≤ m_b (REB) ≤ 4.8
$\log M_0 = 3.0\ m_b$ (REB) + 9.45	m_b (REB) ≥ 4.8

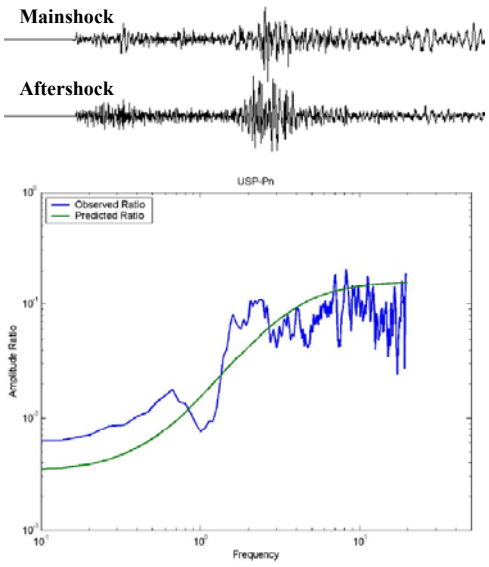


Figure 7. Spectral ratio of a mainshock/aftershock pair in southern Asia, showing good agreement with the cube-root model (green).

Signal Processing Experiments

Controlled experiments with the embedded data offer several opportunities for evaluating algorithms for station signal processing with regard to signal detection as well as estimation of signal parameters. We give examples below with results from initial seismic signal processing experiments. Test data sets were constructed from background noise and scaled signals for three stations: CMAR, FINES, NIL. Signals scaled to magnitudes between m_b =1.8 - 4.5 in increments of 0.1 magnitude units (m.u.) from three large and similar (m_b ~ 5.5) underground nuclear explosions at the Lop Nor test site were embedded in noise of different types – whole background, clean, and reversed clean. The test data sets were processed with DFX signal detection and parameter extraction programs with configurations currently employed at the IDC.

Signal Detection Probabilities

Detection probabilities as a function of m_b were calculated as the ratio of the number of detected signals/total number of embedded signals of magnitude m_b. Figure 8 compares the probabilities for the three stations as a function of m_b. The probability curves are in reasonable agreement with a Gaussian cumulative distribution functions with mean values corresponding to the 50% detection probability threshold, and the standard deviation is a measure of noise amplitude variation. The data in Figure 8 thus support the common assumption of network detection simulations that incremental detection probabilities as a function of m_b can be approximated by cumulative Gaussian distribution functions (Kvaerna and Ringdal, 1999).

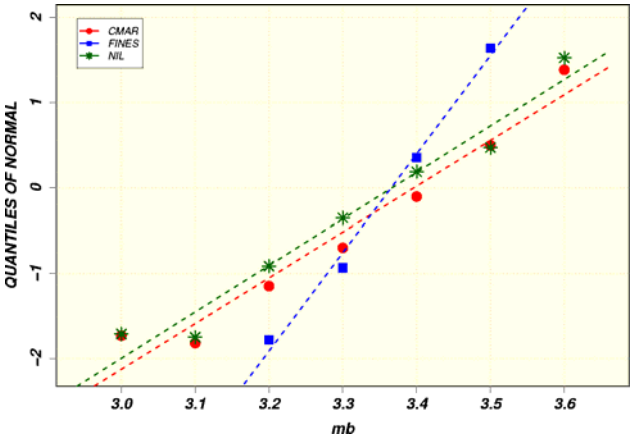


Figure 8. Detection probabilities as a function of m_b for the stations CMAR, FINES, and NIL. A Gaussian scaling is used for the detection probabilities on the vertical axis so that true Gaussian data would follow a straight line. The data points for the detection probabilities follow the fitted lines closely suggesting a that the detection probability as a function of m_b can be approximated with a cumulative Gaussian distribution.

False alarm rates

Estimates of false alarm rates from the rates of unassociated detections become uncertain when based on observations of real data; such unassociated detections include detections of signals from events of unknown origin as well as detections triggered by signal coda. Clean noise, free of signals, can provide a more robust estimate of the false alarm rate, that is unbiased both due to signals of unknown origin and due to coda detections.

In Table 2 we compare the false alarm rates of the three types of noise samples. The false alarm rates for clean and clean-reversed noise with no embedded signals represent genuine false alarms of the detector configuration, whereas for the real noise data some portion of the detections are from signals of seismic events. The number of detections of the whole background noise data that could be associated (using a 4 second allowance for initial P) with seismic events in the ISC database was small (< 10%). The ISC reported close to 200 worldwide events/day for the time period analyzed here, but its event catalog is incomplete for small events. Furthermore, secondary phases or coda detections were not considered in the association, so total detection rates, with ISC associations subtracted, still overestimate the actual false alarm rate. The difference is most striking for NIL, for which the unassociated rate of the clean noise is about a factor of eight lower than that for the real data, while this factor is around 5 for FINES and a little more than 2 for CMAR.

Table 2. Daily Unassociated Detection Rates for Different Noise Types

Station	Clean[1]	Reversed Clean	Whole Background
CMAR	190 | 37	160 | 29	411
FINES	20 | 6	30 | 6	165
NIL	62 | 7	85 | 3	507

1. The two numbers represent the number of unassociated detections in clean noise with no embedded signals/number of additional unassociated detections when signals were embedded.

With signals embedded in the clean noise the false alarm or unassociated rate goes up slightly due to additional coda detections that normally would have been assumed to be false alarms. The rates of such additional unassociated detections are also given in Table 2. The median of this increase in false alarm or unassociated rates when signals are embedded for the clean and reversed clean noise is about 20%.

ROC curves

The performance of a signal detector is often defined by the so-called Receiver Operating Characteristic, ROC, which describe the trade-off between signal detection probability and false alarm probability (Van Trees, 1968). Using scaled signals embedded in clean noise affords the opportunity to construct ROC curves, which is difficult, if not impossible, to do from real data. An example of an ROC curve constructed from detection experiments with the array FINES is shown in Figure 9. The histograms to the left show the distributions of the logarithm of the short term/long term averages (sta/lta), or SNR, for noise (red) and for signals (plus noise) (in blue) of the same size (m_b=3.30). Note that the histograms for the signals include data for signals that were not detected. SNR values for undetected signals could be calculated because of their known embedding times. Gaussian curves draped on the histograms show reasonable agreement with the empirical SNR distributions for both noise and signal data. The detector triggers if the SNR is above a preset threshold and the default threshold used in the processing experiment is marked as a vertical dashed line (green) to the right. A significant portion of the SNR distribution for the explosion signals is below the threshold and hence went undetected. If the threshold were lowered to, e.g., the mean of the SNR distribution (black dashed line to the left marked "NEW THRESHOLD"), the probability to detect an explosion signal would increase significantly (to 50%) without significantly increasing the probability of triggering on noise, or of a false alarm.

In the right diagram the two Gaussian distributions for noise and signal detection probabilities were combined to a standard ROC curve. The “default” and the "new" thresholds are marked showing that by lowering the threshold the false alarm probability would not change much, whereas the detection probability would go up from less than 20% to 50%. It should be noted that the data in Figure 9 represent beam forming with steering of a single beam that is optimum for FINES and the Lop Nor test site.

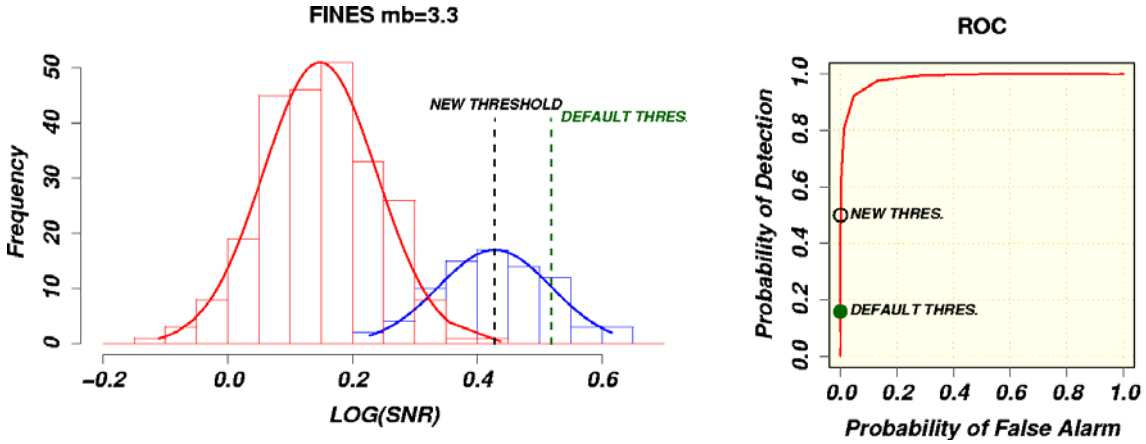


Figure 9. The histograms in the left diagrams for SNR (log scale) of noise (red) and of signals (blue) are used to construct the ROC curve in the right diagram. The default threshold used in the detection experiment is marked as a green dashed line in the left diagram and as a filled dot in the ROC curve. The ROC curve shows that lowering the default threshold (marked as “NEW THRES” in the diagrams) has little effect on the probability of false alarm, but would increase the probability of detection from less than 20% to 50%.

Estimation of Signal Parameters

Apart from detection the processing algorithm that we use (DFX) estimates signal parameters that characterize a detection, such as time of arrival, SNR, amplitude, and for arrays, slowness vector. With the ground truth of the characteristics of embedded signals, the distributions of errors in parameter estimates can be estimated with accuracy. Figure 10 summarizes some statistics of estimation errors in amplitude/period ratios (to the left) and

azimuths of slowness vectors (to the right) for the DFX algorithm for experiments with FINES. The boxplot to the left shows that the automatically estimated amplitude/period ratio becomes increasingly positively biased with decreasing SNR of the detected signal. The bias at the threshold of the detector (log SNR=0.5) is, on average, about 0.3 m.u. The bias is probably caused by low frequency noise that the automatic algorithm cannot account for as the amplitude/period ratios for the downscaled signals were unbiased prior to embedding in the noise (see Figure 5). The bias raises the question whether amplitude/period measurements at low SNR should be used uncorrected for magnitude estimation.

The errors in azimuth estimates (right diagram in Figure 10) grow fast with decreasing SNR. A bias is not as clearly defined as for the amplitude/period ratios, but the scatter increases drastically as indicated by the widening of the boxes which represent 50% of the data around the median. The dashed red lines outline the 50% limits of the azimuth uncertainty (measurement errors) assigned by the automatic algorithm, which clearly underestimates the spread of the actual azimuth errors at low SNRs around 4 and below. An analysis of the empirical distributions of the amplitude and azimuths errors revealed that they are well represented by a Gaussian for all SNRs.

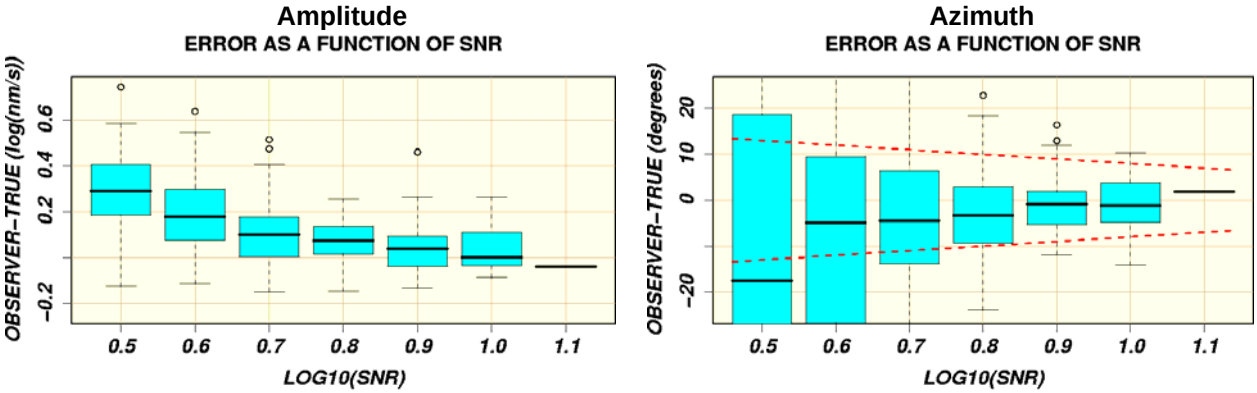
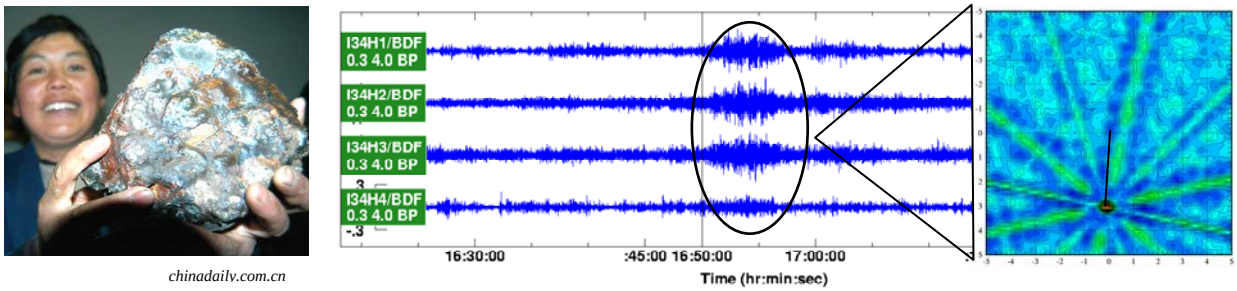


Figure 10. Errors in amplitude/period ratios (left) and azimuth (right) of the automatic DFX algorithm for signals embedded in clean noise of the array FINES. The boxplots show the errors as a function of log SNR; the dashed lines for azimuth in the boxplot to the right outline the 50% limits of errors assigned by the automatic algorithm (measurement errors).

Seismo-Acoustic Event Database and Scaling

Detection of infrasound signals from seismo-acoustic events (e.g., mine blasts) can provide valuable insights into a variety of research topics, such as infrasound propagation, seismic vs. acoustic coupling and the effects of wind-generated noise on infrasound detection. It is our objective to provide to the nuclear monitoring research community a controlled data set where scaled infrasound signals are embedded in a variety of noise conditions. This will be particularly useful for evaluations of detection algorithms. Currently evaluations with respect to variations in ambient noise are dependent on finding data sets where signals from known sources are detected in a variety of noise conditions. Given the current dearth of such infrasound ground truth, a comprehensive evaluation of infrasound detection capabilities using actual recordings is difficult.

We are following the same basic methodology for infrasound that we used for seismic data (described above). Namely, we are building a background noise database from historical recordings, collecting high-SNR infrasound signals from known sources, applying a source scaling function to the recorded data, and embedding the scaled signals in the background noise. In contrast to nuclear explosion or earthquake source theory, general theoretical models cannot be used to predict the scaling of infrasound signals from mine blasts. This is in large measure due to the fact that differing mine practices, acoustic coupling and other local conditions vary immensely from mine to mine and region to region. Therefore we are following an empirical approach to assess the scaling of infrasound signals and limiting the scope of the scaled and embedded data. We are building the data set such that each scaled/embedded event represents effectively the same source and meteorological conditions as the original event, with only the signal amplitude (by no more than one order of magnitude) and the ambient noise conditions varying.



chinadailv.com.cn

Figure 11. Bolide near Lanzhou, China on December 11, 2004 recorded on the IMS infrasound array I34MN and reported in the local media.

We obtained seismo-acoustic recordings (seismic and infrasonic signals from the same source) from 23 mine blasts (c) and 1 bolide (Figure 11) in central Asia for inclusion in our signal library. The mine-blast data offer an opportunity to look for evidence of non-linear source scaling in the infrasound data. b shows a section of 9 infrasound signals recorded at I31KZ for mines at ranges between 50 and 400 km. Traces 1 and 3 originating from the same mine, however, had signal amplitudes differing by a factor of about 10. Similarly traces 2 and 4, again originating from the same mine, had signal amplitudes differing by a factor of about 3. In both cases, spectra ratio comparisons showed that the signals scaled approximately linearly with frequency, consistent with the visual similarity of the waveforms. Thus, based on this admittedly limited data set we are proceeding with a linear scaling function and embedding these in a variety of noise conditions.

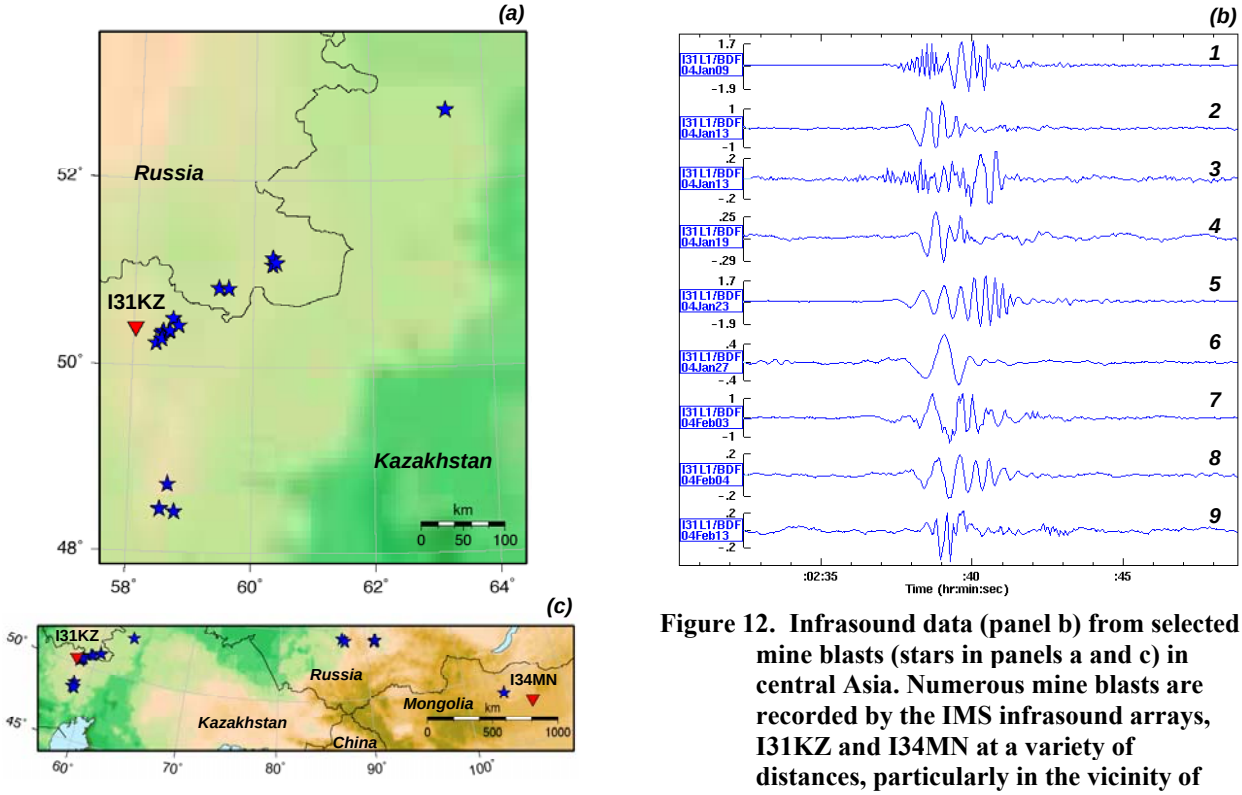


Figure 12. Infrasound data (panel b) from selected mine blasts (stars in panels a and c) in central Asia. Numerous mine blasts are recorded by the IMS infrasound arrays, I31KZ and I34MN at a variety of distances, particularly in the vicinity of I31KZ (panel a).

CONCLUSIONS AND RECOMMENDATIONS

We developed a network data set where scaled signals were embedded in varying background noise conditions, including carefully constructed clean noise. Analysis of reprocessed scaled data show that both a Mueller-Murphy model for nuclear explosions, and modified Brune model with cube-root corner frequency scaling and an empirical $\log M_0$-m_b relation for earthquakes, give consistent amplitude scaling for events in southern Asia.

The detection experiments we executed illustrate some of the benefits that using an embedded data set can provide for assessment of signal detection and characterization algorithms. We found clear evidence for an amplitude bias for low SNR signals and that measurement uncertainties for azimuth probably underestimate the error at low SNR. In general we conclude that experiments with an embedded data set can:

- Validate assumptions for detection probabilities underlying network simulations of real data
- Provide accurate estimates of station detection probability as a function of source strength
- Provide unbiased estimates of false alarm rates
- Accurately map ROC curves for relative assessment of signal detectors and detector configurations
- Estimate the statistical characteristics of errors of signal parameters such as arrival times, slowness vectors, and signal amplitudes.

The background noise, scaled signals, embedded waveforms and relevant meta-data from this effort are available to the monitoring research community via the RDSS web site (http://www.rdss.info/, Woodward et al. 2005).

REFERENCES

Brune, J. N. (1970), Tectonic stress and spectra of seismic shear waves from earthquakes, *J. Geophys. Res.* 75: 4997-5009.

Engdahl, E.R., R. van der Hilst & R. Buland (1998), Global Teleseismic Earthquake Relocation with improved Travel Times and Procedures for Depth Determination, *Bull. Seism. Soc. Am.*, 88: 722-743.

Hanks, T. C., and W. H. Bakun (2002). A bilinear source-scaling model for M-logA observations of continental earthquakes, *Bull. Seism. Soc. Am.* 92: 1841-1846.

Kohl, B., T. Bennett, I. Bondár, B. Barker, W. Nagy, C. Reasoner, and J. Hanson (2004), Development of a Network Data Set for Evaluating Detection and Network Processing Performance, in *Proceedings of the 26th Seismic Research Review: Trends in Nuclear Explosion Monitoring*, LA-UR-04-5801: Vol. 2, pp. 725-734.

Kværna, T. and F. Ringdal (1999), Seismic Threshold Monitoring for Continuous Assessment of Global Detection Probability, *Bull. Seism. Soc. Am.*, 89: 946-959.

Mayeda, K., and W. R. Walter (1996), Moment, energy, stress drop, and source spectra of western United States earthquakes from regional coda envelopes, *J. Geophys. Res.* 101: 11,195-11,208.

Mueller, R. A., and J. R. Murphy (1971), Seismic characteristics of underground nuclear detonations, part 1: seismic spectrum scaling, *Bull. Seism. Soc. Am.* 61: 1675-1692.

Murphy, J. R. (1977), Seismic source functions and magnitude determinations for underground nuclear detonations, *Bull. Seism. Soc. Am.* 67: 135-158.

Patton, H. J. (2001), Regional magnitude scaling, transportability, and MS:mb discrimination at small magnitudes, *PAGEOPH* 158: 1951-2015.

Taylor, S. R., A. A. Velasco, H. E. Hartse, W. S. Phillips, W. R. Walter, and A. J. Rodgers (2002), Amplitude corrections for regional seismic discriminants, *PAGEOPH* 159: 623-650.

Van Trees, H. L., (1968), *Detection, Estimation and Modulation Theory*, John Wiley and Sons.

vonSeggern, D. H. and Blandford, R. (1972), Source time functions and spectra for underground nuclear explosions, *Geophys. J.* 31: 8-97.

Woodward, R., M. Skov, M. Bahavar, G. Davis and Y.-L. Kung (2005), Data and tools to support nuclear explosion monitoring research and development, in *Proceedings of the 27th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies,* in current Proceedings.

INTEGRATED SEISMIC EVENT DETECTION AND LOCATION BY ADVANCED ARRAY PROCESSING

T. Kvaerna[1], S. J. Gibbons[1], F. Ringdal[1], and D. B. Harris[2]

NORSAR[1] and Lawrence Livermore National Laboratory[2]

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract Nos. DE-FC52-03NA99517[1] and W-7405-ENG-48[2]

ABSTRACT

We have developed a prototype system for the automatic monitoring of seismic events from sources of interest using regional seismic arrays. The aim of such a system is to provide significantly improved location estimates for low-magnitude events compared with current automatic approaches, combined with a low false alarm rate. The system is a generalization of an algorithm developed under a pilot project to monitor events from the Kovdor mine in NW Russia using only the ARCES regional array at a distance of 300 km, applying carefully calibrated processing parameters based upon previous observations of confirmed events at the site of interest. The new automatic system is therefore suited, but not restricted, to the single array case. We have applied the process to the Fennoscandian arrays and the arrays in Kazakhstan.

As an initial step, every detection at each of the stations employed is reprocessed in two stages: firstly, the onset time is re-estimated using an autoregressive method and, secondly, the slowness is estimated using broadband f-k analysis in several predetermined fixed frequency bands. The slowness observed can vary considerably from one frequency band to another and the frequency band providing the most stable estimates for repeated observations from a given source varies greatly from site to site. For the local and regional events considered, frequencies below 2 Hz rarely provide useful slowness estimates for the Fennoscandian arrays due to the strong background noise. Frequencies above 4 Hz were not used in the reprocessing of the Kazakhstan data due to signal incoherence over the arrays.

The automatic monitoring system is based upon two types of templates: a site template that lists which phases are anticipated at which stations at which times, and a template for each phase specifying a permissible range of slowness and azimuth observations. In order to calibrate the templates required to identify phase arrivals from events at a given site, a dataset of confirmed events is required. For the mining regions on the Kola Peninsula of NW Russia, lists of confirmed industrial blasts were obtained for many different mines, allowing an extensive study of variability of slowness estimates in various frequency bands for each anticipated phase. Such ground truth information was not available for the Swedish mining regions or industrial explosions in Kazakhstan. However, given a small number of events known to have occurred at different sites, lists of events guaranteed to have occurred in the near vicinity of these master events have been generated by performing waveform correlation on signals from likely candidate events.

Conventional f-k analysis and beamforming assume a plane wavefront which is coherent across the array; this assumption breaks down due to refraction and scattering, leading to energy loss in beamforming and bias in slowness estimates. In matched field processing, the plane-wave steering vectors are replaced with empirical steering vectors estimated from observations of phases from events in the region to be monitored. An efficient suite of calibration software has been developed in this project to filter suites of waveforms from training event sets into a large number of narrow bands and to estimate matched field steering vectors for each band. We present an example whereby a set of steering vectors was calibrated for the Pn-phase at ARCES from compact underground explosions at the Kirovsk mine. In the 7.8-12.5 Hz frequency band, the matched field beam captures a factor of 2 more energy than the conventional beam and, when filtered above 10 Hz, the factor is closer to 3.

OBJECTIVE

This two year collaboration between NORSAR and Lawrence Livermore National Laboratory (LLNL) has explored improvements to the automatic detection and location of seismic events using regional arrays. At the heart of the study has been the calibration of processing parameters for the detection and location of events from a specific region using observations of previous ground truth events at the sites of interest. The goal is to attribute, with a high degree of confidence, automatically located events to active mines or areas with known recurring seismicity. The study has examined sites in Fennoscandia and Kazakhstan using the seismic arrays in these regions.

The signals at a given array station, resulting from a set of events from a site with recurring seismicity, are likely to display common characteristics which may be exploited in order to identify subsequent events from the same region. A template describing the measurements which can be anticipated at a given station at a given time can be used to judge whether or not a detected signal is the likely result of an event from the site of interest. Such templates must be calibrated by investigating the variability of measurements made from events confirmed to have taken place at the sites; such calibrations have been the main focus of this investigation. We have, in addition, explored the potential of applying advanced new "matched field" array processing methods in order to compensate for array processing loss due to refraction and scattering, thus enhancing array gain at high frequencies.

RESEARCH ACCOMPLISHED

Introduction

The ARCES regional array in northern Norway is a primary International Monitoring Station (IMS) seismic station within a few hundred kilometers of many active mining regions, both in north-west Russia and northern Sweden (Figure 1). Signals from routine industrial explosions at these sites in fact dominate the ARCES detection lists and their identification and location require considerable analyst time. Fully automatic event locations at NORSAR are currently provided using the Generalized Beamforming (GBF: Kværna and Ringdal, 1989) system which associates detected phases from the entire network of regional arrays. The collection of ground truth data from mining explosions on the Kola Peninsula (under the DOE funded contract "Ground-Truth Collection for Mining Explosions in Northern Fennoscandia and Russia"; Harris et al., 2003) has provided an excellent opportunity to assess the current state of the automatic event detection and location procedures and to examine approaches for improving it.

The ground truth information for the mining regions on the Kola Peninsula provides the origin times of shots from 13 mines from the four distinct Russian mine clusters color-coded in the upper panel of Figure 1. The corresponding colored symbols in the lower panel of Figure 1 indicate the fully automatic locations from the GBF system for the confirmed events from these mines between October 1, 2001, and September 30, 2002. It is immediately apparent that the distances between the automatic location estimates and the true locations vary enormously. In particular, given a single one of these automatic locations for a recently detected event, it is impossible to ascribe the signal to any of the sources shown without performing a full time-consuming manual analysis upon the signal. The most significant reasons for the large variance of location estimates from the GBF system are the following:

- Azimuth and slowness estimates are performed on data filtered in a frequency band which varies from detection to detection. The frequency band is set in order to optimize the signal-to-noise ratio (SNR) and it is demonstrated in Kværna et al. (2004), the extent to which the slowness estimates for the same set of events become more stable when estimated in fixed frequency bands.
- Many of the events have complicated firing sequences which can lead to an incorrect association of phases. For instance, if two similar blasts follow within seconds of each other, it is possible that an S-phase from the second shot may be associated with the P-phase from the first shot leading to a location estimate at too great a distance. Other combinations of incorrect coda phase associations can lead to similar spurious location estimates.
- The GBF system follows somewhat empirically determined rules which help to determine which of several candidate sources (hypocenter and origin time) is the most likely to have produced a given set of phases. A single trial hypocenter which corresponds to a large number of phases may score more highly than, for example, two hypotheses for different events which would give a more accurate description of the cause of the detected phases.

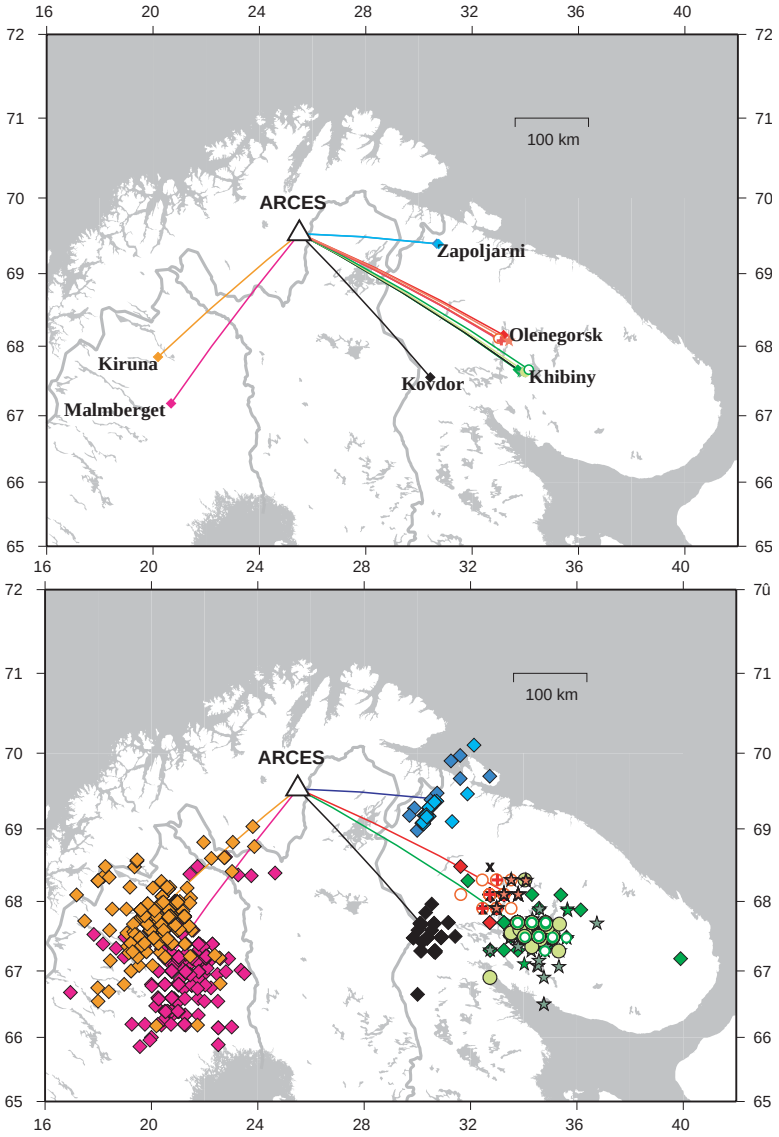


Figure 1. Location of the ARCES regional array in relation to mining regions in northern Sweden and on the Kola Peninsula, Russia (top panel) and automatic event locations using the NORSAR GBF system for events known to have occurred at the sites indicated (lower panel).

Under a pilot project, a fully automatic system for the identification and location of events occurring at, and in the close vicinity of, the Kovdor mine in north west Russia, using only the ARCES regional array, was developed. The details of this procedure are provided by Gibbons, Kværna, and Ringdal (2005). Fundamental to this system is the consistency provided by slowness measurements in a fixed frequency band of a given phase from a given site. If, for a given detection, a slowness estimated from broadband f-k analysis in a fixed frequency band falls within the narrow confines of a template calibrated from the observations of previous events, we can immediately form a hypothesis that an event at our monitored site occurred at the indicated time. We subsequently test this hypothesis by examining whether or not slowness measurements in time-windows fixed relative to the hypothetical origin time are consistent with the existing body of observations from that site. Gibbons, Kværna and Ringdal (2005) demonstrate that the single array automatic location estimates are a significant improvement on the existing automatic solutions and are comparable to multi-array analyst locations. The greatest difficulty encountered in this study was the problem in the identification of secondary phases (usually the result of complicated source time histories) which led to many events which could not be located in this manner; these events had to be filed for analyst review.

The goal of the current project has been to construct an integrated framework in which the philosophy behind the single array monitoring system of Gibbons, Kværna and Ringdal (2005) could be generalized to a range of different source regions and seismic arrays. The ground truth collection project had also acquired information on a very large number of events from the Zapoljarni, Olenegorsk, and Khibiny mining regions (Figure 1). These mining clusters present an additional complication in that several mines operate within short distances of each other; the inter-mine separation is not generally large enough for a single array at a distance of several hundred kilometers to be able to differentiate between two sites from a slowness measurement alone, but it is sufficiently large for the fixed time-window scheme of the Kovdor monitoring process to be compromised.

For other sites, such as the Kiruna and Malmberget mining regions in northern Sweden, there exists the problem that we do not possess ground truth data on routine explosions which are necessary in order to calibrate the templates. The approximate times of the daily routine detonations were however known for both sites which made it possible to identify very likely events and perform a careful analyst location for each of these. With a few master events for each mine, large numbers of events were subsequently identified using a correlation detection algorithm as being almost certain to have originated within a few hundred meters of the master events. The detected events which corresponded to reasonably high correlation coefficients and which displayed a satisfactory signal-to-noise ratio were selected to build up the databases of events assumed to originate from these source regions. All the orange and magenta symbols in the lower panel of Figure 1 were identified in this way.

Calibrating the event and phase templates

Given a sufficiently large set of events from the sites for calibration, it is necessary to identify properties of the resulting wavefields which provide the most stable characteristics for the subsequent identification of new events. The most stable property is almost always the slowness estimate for the initial P-arrival from each event; this is demonstrated clearly in Figure 8 of Kværna et al. (2004). It also emerges that the frequency band which provides the most stable slowness estimates for a given event population varies greatly (c.f. Figure 6 of Kværna et al., 2004). For the Zapoljarni mines, for example, the 2-4 Hz frequency band gave a far smaller spread of slowness estimates than other frequency bands whereas, for the Khibiny mines, the 4-8 Hz frequency band gave the most consistent estimates.

This raises the question of how a "likely candidate phase" should best be identified and, given the results from the processing of the Kola ground truth mining events, it was deemed that the best procedure would be to perform broadband f-k analysis in a wide range of fixed frequency bands for every detection made by the arrays. In this way, a detection list could simply be scanned by a number of "trigger templates" each examining the frequency band (or set of frequency bands) which provided the best slowness estimates for the target phase. It would be sensible to construct a trigger template for an initial P-arrival from a Zapoljarni event that would be activated when a slowness estimate in the 2-4 Hz band fell within the permitted bounds (obtained from the set of training events) and a trigger template for a P-phase from a Khibiny event which would test the slowness in a higher frequency band. Experience has however shown that it is advisable to provide a few alternative trigger combinations since the optimal frequency band may give a spurious slowness estimate as a result of, for example, an interfering signal. It was found to be helpful to generate panels displaying the slowness estimates and corresponding beams for each of the fixed frequency bands to allow an "at-a-glance" assessment of the quality of each detection. An example of such a panel is displayed in Figure 5 for a regional phase arrival at one of the arrays in Kazakhstan.

It was also demonstrated by Kværna et al. (2004) that autoregressive onset time estimates frequently provided far better estimates than amplitude-only-based methods and so a two-stage reprocessing system was activated for each array in which we first obtain the best possible arrival time estimate and then obtain the slowness estimates in each of the specified frequency bands. This reprocessing procedure can be performed for an arbitrarily specified time and can consequently be applied to any form of detection. It is currently applied for every conventional detection from each of the regional arrays which supply data to NORSAR. However, it could conceivably also be applied to correlation detections, detections from matched-field processing, or simply at times for which there is reason to suspect that an arrival of interest may have occurred.

On each occasion that a trigger template is activated, an event hypothesis is formed for the site region for which the activated template is calibrated. Each such calibrated site has a corresponding "site template" listing the phases which ought to be observed whenever an event at that site occurs. Each of these phases corresponds to a "phase template"

stating at which time the phase should be observed, which range of slowness values are consistent with the phase in a specified set of frequency bands, how large an SNR should be observed to support the assumption that the phase has indeed arrived at the stated time, and which range of autoregressive arrival time estimates are acceptable under the considered event hypothesis. The following section considers an automatic event detection algorithm using the very simplest form available for site templates: a single P-phase and a single S-phase recorded at a single array.

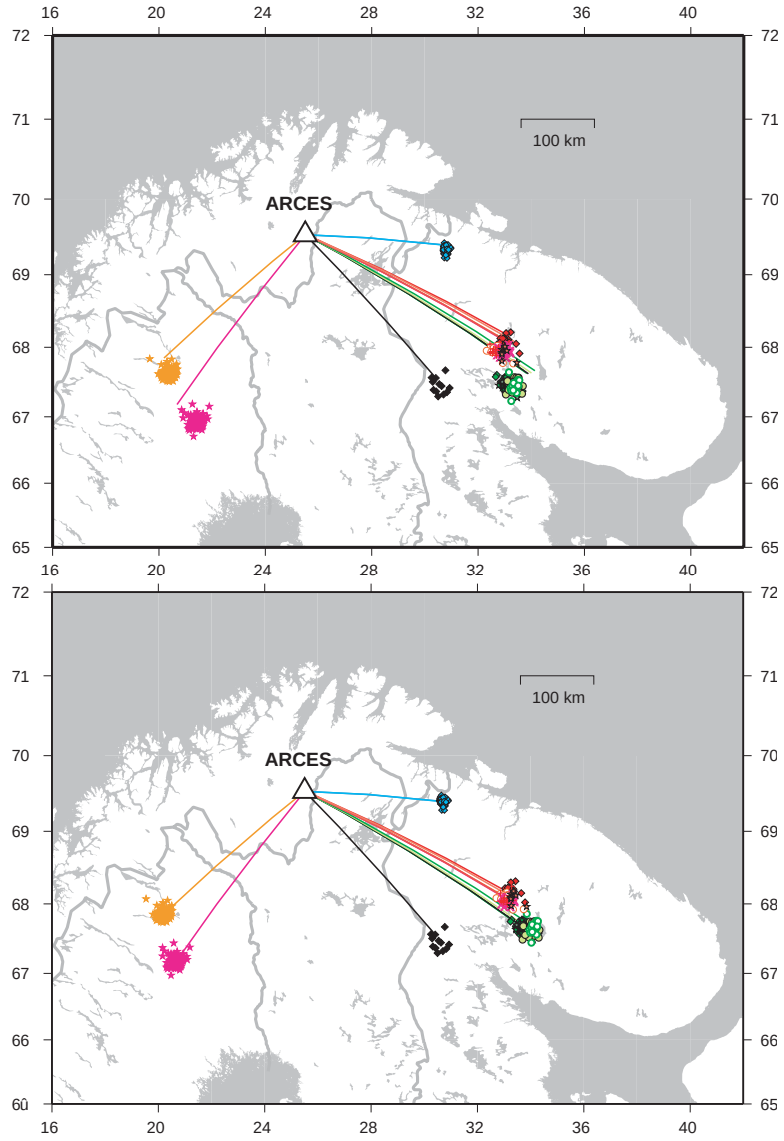


Figure 2. Locations of the events in the lower panel of Figure 1 based upon the site-specific location algorithms with slowness and azimuth estimated in fixed frequency bands selected for each site, and phase onset times estimated by autoregressive methods. The upper panel features the locations made without systematic slowness and travel-time corrections in the location procedure and the lower panel features locations made applying corrections. Note that not all of the events shown in Figure 1 are included in these figures since events which fail to match required components of the templates are excluded prior to the location procedure. Each event which is located showed sufficient evidence of belonging to the geographical region represented by the template; this essentially eliminates the possibility of events being located at large distances from the target sites. Locations are made using the HYPOSAT algorithm (Schweitzer, 2001) using only the Pn and Sn phase arrivals at the ARCES array (Pg and Sg for the Zapoljarni mines).

A template-based event location procedure

For each of the mining regions displayed in Figure 1, a site template was formulated listing two anticipated phases: the initial P-arrival at ARCES (which would be used to trigger an event hypothesis) and the first S-arrival (only used to confirm an event hypothesis). The first S-arrival at ARCES is Sn for all of the sites shown except for the Zapoljarni mines for which the Sg-phase is the first secondary phase to arrive. The slowness bounds for each of the phase templates were set according to the variability observed for the sets of training events (see Kværna et al., 2004). Since an initial P-arrival at ARCES from one of the Khibiny mines would not permit the system to exclude the possibility that the phase originated from one of the other Khibiny mines, it was decided that a site-template would be set up for each one of the mines, and all would trigger for an initial P-phase from any one of the Khibiny sites. Each of the site templates would differ in the initial setting of the time-window for the examination of the secondary phases but, unlike the Kovdor monitoring process (Gibbons, Kværna and Ringdal, 2005) in which slowness estimates were only performed in time-windows fixed relative to the initial P-arrival, we allow here a small deviation through a limited number of iterations. In practice, this means that we measure the slowness in a time-window fixed relative to the initial P-arrival, then form a beam steered by these parameters, measure a new autoregressive onset time, then confirm that the new time falls within a permissible time-window and that the slowness measured at the new time falls within the accepted range. The new slowness is subsequently used to form a new beam and the procedure is repeated. If, at any stage, a test is failed then the event hypothesis is rejected (or at least filed for analyst review); this will limit how far from the specified site template the solution can deviate but should permit a location to be found even if the event is a short distance away from the exact site for which the site-template was tuned.

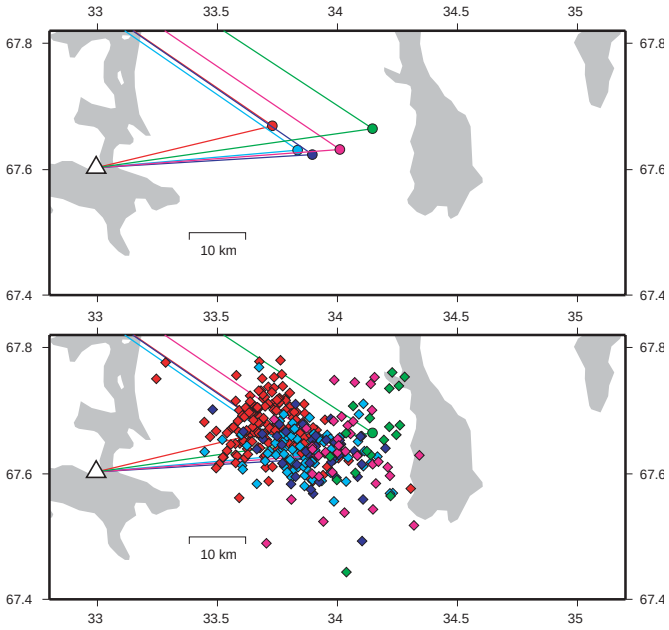


Figure 3. Zoom-in view of the locations of the mines of the Khibiny Massif near the town of Apatity in NW Russia (upper panel) together with automatic, single-array locations of confirmed events from these mines between October 2001 and September 2002 (lower panel). All locations are made using HYPOSAT using only the Pn and Sn phases only from the ARCES array at a distance of approximately 410 km to the North West. The triangle indicates the location of the Apatity array (not used in these locations) and the colored lines indicate the shortest routes from the mines to the seismic arrays.

Figure 2 shows the automatic event locations obtained for the events shown in the lower panel of Figure 1 using the slowness and azimuth estimates obtained from the frequency bands specified in each of the phase templates and phase arrival times measured by the autoregressive onset estimates. This figure indicates the vast improvement to the location estimates afforded by this controlled process and also the importance of applying slowness and travel-time corrections to the solution prior to calling a location algorithm. If the azimuth measured in the most stable frequency band is interpreted directly as being the geographical backazimuth, large systematic biases (of up to 50 km) can be

introduced. In fact, the Kovdor mine is unique among these mining regions in that the systematic azimuthal bias for this site can essentially be ignored; the systematic azimuth bias for the other sites is often of the order of 5 degrees.

Not every event from the mining regions was able to be located using this single-array, two phase site template algorithm. Many events were excluded for having failed a test (a slowness estimate which was not consistent with a phase template or an arrival time estimate which did not fall within the permissible range or which displayed too low an SNR). In particular, many events from the Kovdor mine failed this two-phase algorithm due to the failure of the Sn phase to record an acceptable slowness value in the pertinent time-window; this is often caused by a low amplitude, emergent Sn phase compounded by complicated firing sequence. Gibbons, Kværna, and Ringdal (2005) improve the number of events located by also considering the far stronger Lg phase. An alternative strategy would be to attempt to detect phase arrivals at a different station, which would reduce the importance of detecting a secondary phase.

Figure 3 shows a zoom-in of the calibrated single-array locations for the mines on the Khibiny massif. While the events from the Kirovsk mine (red symbols) cluster to the West and events from the Norpakh mine (green symbols) cluster to the east, as would be hoped from the source locations, we see that this single array process is unable to resolve these populations. This limit of location resolution is consistent with that experienced by Gibbons, Kværna, and Ringdal (2005) for the Kovdor mine. However, even without use of the nearby Apatity array, this location algorithm is probably sufficiently good to provide a preliminary clustering of events which can then undergo a full waveform source identification procedure.

Application of calibrated array processing to data from Kazakhstan

Like Fennoscandia, Kazakhstan is a region containing several seismic arrays and large numbers of routine industrial explosions. Figure 4 shows the locations of the arrays together with reviewed bulletin locations of events over a one month period for which at least one regional phase was recorded at the Akbylak array. It is clear from this distribution that a large number of events cluster around small regions which coincide with the locations of mines.

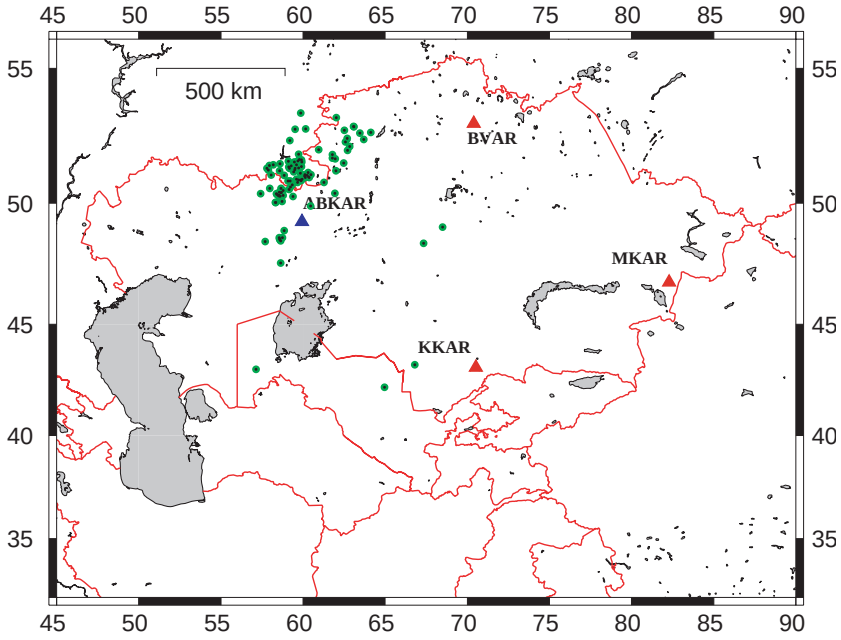


Figure 4. Locations of the four 9-element arrays in Kazakhstan together with locations from the Kazakhstan NDC reviewed event bulletin of all events during May 2004 for which regional phases were recorded at the Akbylak array (ABKAR). Also shown are the arrays at Borovoye (BVAR), Makanchi (MKAR), and Karatau (KKAR).

While ground truth information for the mining sites is not yet available, it appears that the waveform-correlation bootstrapping approach applied to the Swedish mining regions may also be applied here. Many of the analyst-located events in the Kazakhstan bulletin occur within a small geographical region at times at which the mines are known to conduct routine explosions. When the waveforms are filtered in a sufficiently low-frequency band, many events are observed to show a high degree of waveform semblance indicating a small separation between the source regions.

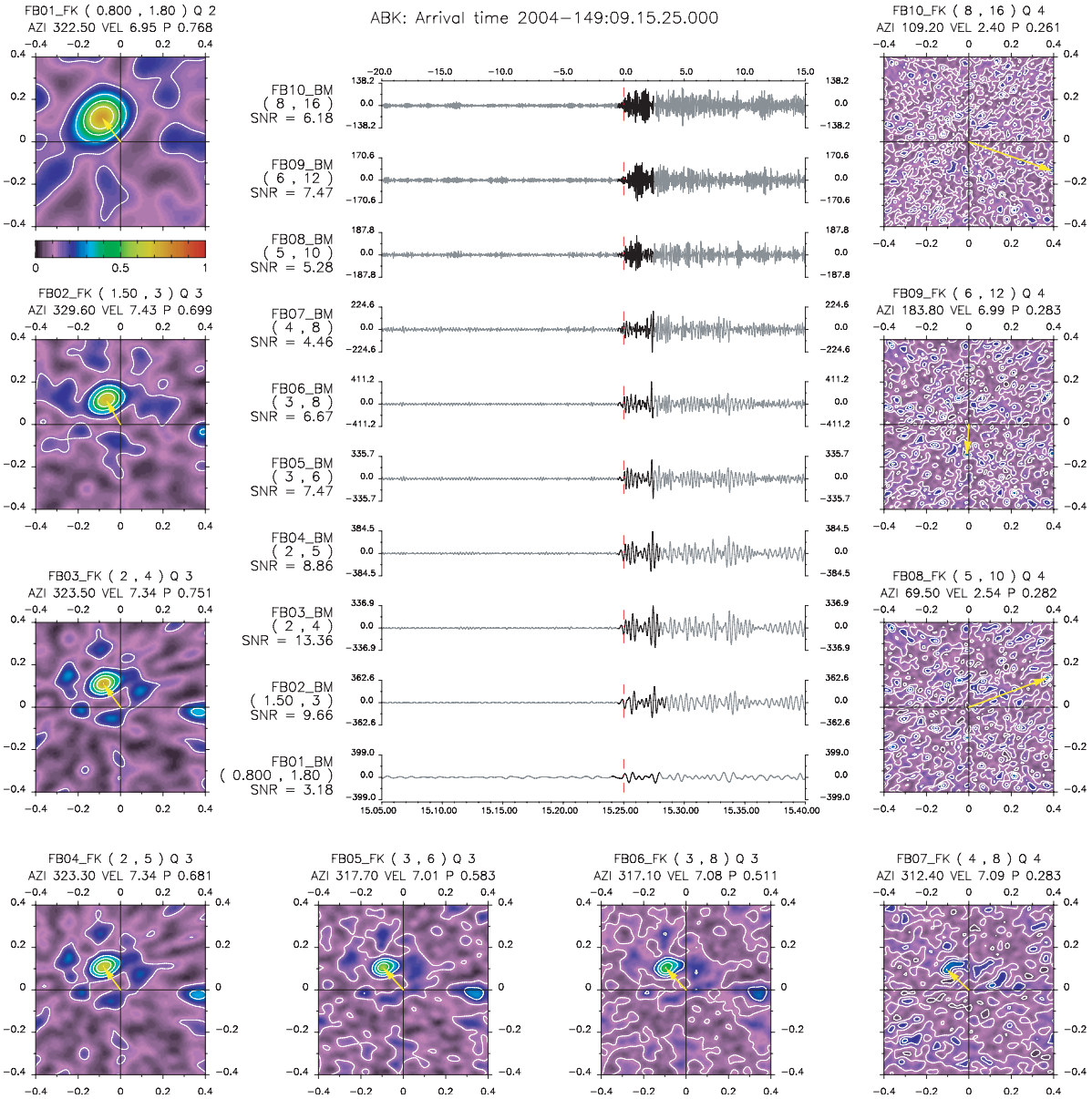


Figure 5. Slowness and azimuth estimates in a predetermined set of fixed frequency bands for a first P-arrival from regional event at a distance of approximately 150 km from the Akbylak array in Kazakhstan.

It appears that the events in the clusters identified so far exhibit a similar "slowness and azimuth as a function of frequency band" to that observed at ARCES (see Figure 5). However, the frequency bands in which regional signals are best observed are very different at the arrays in Kazakhstan from those in Fennoscandia. For ARCES and the other Fennoscandian arrays, frequencies below 2 Hz are essentially unusable for regional signals because of the high-microseismic noise. The SNR for most of the mining explosions simply improves as the frequency increases and

the optimal analysis frequency band is a trade-off between SNR and coherence. In Kazakhstan, the low frequencies do allow good observations of regional phases and the frequency bands above 4 Hz for these arrays do not give good slowness estimates due to signal incoherence. The set of fixed frequency bands for the reprocessing is therefore chosen accordingly.

Compensation for array loss using matched-field processing

Classical beamforming transforms the incident waveforms using steering vectors which are set according to the assumption of a plane-wavefront. Array loss occurs when the coherence of the waveforms is diminished or the plane-wavefront assumption is otherwise violated due to diffraction or scattering. It has already been noted that many of the signals of interest exhibit the best (single channel) SNR at high frequencies for which beamforming is ineffective due to low waveform semblance. A method of compensating for such array loss by replacing the theoretical steering vectors with empirical steering vectors calibrated from measurements of the wavefield structure, the so-called "matched-field processing" method, has already been employed successfully in the field of underwater acoustics (Baggeroer et al., 1993). The steering vectors are calculated as the eigenfunctions of sample covariance matrices obtained from narrow-band filtered waveforms from populations of events known to have come the same source location.

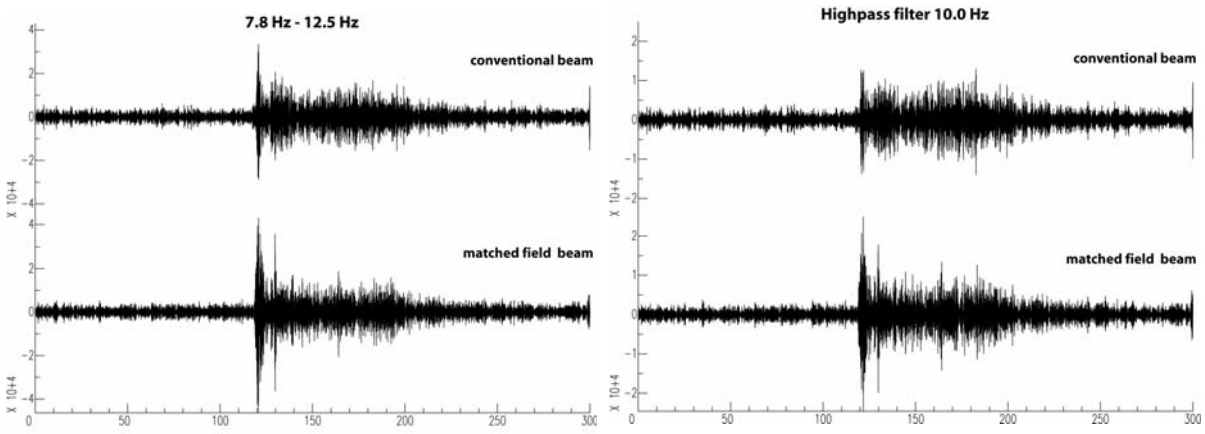


Figure 6. A comparison of conventional and emprical matched field Pn beams using the ARCES array for signals generated by a confirmed compact underground explosion at the Kirovsk mine on the Khibiny massif. A low-frequency array geometry was used here to illustrate the possible gains of matched field processing in situations where signals are only partially coherent. The matched field beam captures between and 3 and 5 dB more Pn energy depending on the frequency band considered.

These calibrations can be applied by performing narrow-band filtering upon an incoming data-stream and beamforming the resulting waveforms using the empirical steering vectors. Figure 6 illustrates the improvement in the SNR on the matched-field beam over a conventional beam for a Pn arrival from a Khibiny massif event. The effect may be even greater for high-frequency regional signals on the arrays in Kazakhstan for which loss of coherence at high frequencies presents a serious problem.

CONCLUSIONS AND RECOMMENDATIONS

We have designed a framework for the automatic monitoring of seismic events from sites of interest using regional seismic arrays. Under the current automatic detection and event location algorithms employed at NORSAR and elsewhere, slowness and azimuth estimates are calculated using broadband f-k analysis in frequency bands which are determined on a detection by detection basis in order to capture the best possible part of the signal. Whereas this generally leads to the best SNR it also leads to a demonstrable variability in automatic event location estimates (Figure 1) which necessitates a costly manual event location procedure for every such signal detection. Under the system proposed here, each detection is reprocessed by performing f-k analysis in each of several fixed frequency bands. This has the advantage that, since the fixed band slowness estimates are typically more stable for events from any given site and that the optimal frequency band varies from site to site, candidate detections can be readily picked

out by a process which detects slowness estimates in a specified band (or combination of bands) which fall within a given set of calibrated bounds. Each time such a candidate phase detection causes a trigger, a full process testing a site template can be initiated which tests any appropriate combination of phases at any appropriate combination of stations. For each time of an anticipated arrival, a new fixed-band slowness estimate is initiated together with an arrival time determination procedure. We have demonstrated that, even in the case of a single array with a two-phase site template, the fully automatic location estimates provide a great improvement over the existing solutions. We have also demonstrated the need to correct for systematic bias in direction and travel time measurements in the location procedure.

The most time consuming and difficult part of this procedure is in the calibration of templates for events known to have originated from a given site. This has been facilitated in this study by the provision of ground truth information from operators of the mines in NW Russia and the application of waveform correlation procedures elsewhere. To conclude this project, we will examine the effectiveness of such a procedure for sites of recurring seismicity in Kazakhstan.

ACKNOWLEDGEMENTS

We are grateful to Vladimir Asming and colleagues at the Kola Regional Seismological Center in Apatity, Russia, for collecting ground truth information from the Russian mine-operating companies as to the sites and times of routine industrial explosions. We would also like to thank staff at the Kazakhstan NDC for providing data from the Kazakhstan array network together with automatic and reviewed event bulletins.

REFERENCES

Baggeroer, A. B., W. A. Kuperman, and P. N. Mikhalevsky (1993), An overview of matched field methods in ocean acoustics, in *IEEE Journal of Oceanic Engineering*, 19(4): 401-424.

Gibbons, S. J., T. Kværna and F. Ringdal (2005), Monitoring of seismic events from a specific source region using a single regional array, in *J. Seismol.* (in press).

Harris, D. B., F. Ringdal, E. O. Kremenetskaya, S. Mykkeltveit, J. Schweitzer, T. F. Hauk, V. E. Asming, D. W. Rock and J. P. Lewis (2003), Ground-truth collection for mining explosions in Northern Fennoscandia and Russia, in *Proceedings of the 25th Seismic Research Review - Nuclear Explosion Monitoring: Building the Knowledge Base*, LA-UR-03-6029, Vol. 1, pp. 54-63.

Kværna, T., S. J. Gibbons, F. Ringdal, and D. B. Harris (2004), Integrated Seismic Event Detection and Location by Advanced Array Processing, in *Proceedings of the 26th Seismic Research Review - Trends in Nuclear Explosion Monitoring*, LA-UR-04-5801, Vol. 2, pp.742-751.

Kværna, T. and F. Ringdal (1989), A multichannel processing approach to real time network detection, phase association and threshold monitoring, in *Bull. Seism. Soc. Am.* 79: 1927–1940.

Schweitzer, J. (2001), HYPOSAT - An enhanced routine to locate seismic events, in *Pure Appl. Geophys*. 158: 277-289.

ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION AND SCIENTIFIC INFORMATION MANAGEMENT

Stanley D. Ruppert, Douglas A. Dodge, Annie B. Elliott, Michael D. Ganzberger, Teresa F. Hauk, and Eric M. Matzel

Lawrence Livermore National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-48

ABSTRACT

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation and scientific data management efforts and discuss frameworks to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research. The software and scientific automation initiatives directly support the rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure and analyze data, and provide a framework for research dataset integration. The automation also improves the researchers ability to assemble quality controlled research products for delivery into the NNSA Knowledge Base (KB). The software and scientific automation tasks provide the robust foundation upon which synergistic and efficient development of GNEM R&E program seismic calibration research may be built.

The task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration product construction are susceptible to automation and future economies. We are applying software and scientific automation to problems within two distinct phases or "tiers" of the seismic calibration process. The first tier involves initial collection of waveform and parameter (bulletin) data that comprise the "raw materials" from which signal travel-time and amplitude correction surfaces are derived and is highly suited for software automation. The second tier in seismic research content development activities includes development of correction surfaces and other calibrations. This second tier is less susceptible to complete automation, as these activities require the judgment of scientists skilled in the interpretation of often highly unpredictable event observations. Even partial automation of this second tier, through development of prototype tools to extract observations and make many thousands of scientific measurements, has significantly increased the efficiency of the scientists who construct and validate integrated calibration surfaces. This achieved gain in efficiency and quality control is likely to continue and even accelerate through continued application of information science and scientific automation.

Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. Whereas it was possible for individual researchers to download individual waveforms and make time-consuming measurements event by event in the past, with the terabytes of data available today, a software automation framework must exist to efficiently populate and deliver quality data to the researcher. This framework must also simultaneously provide the researcher with robust measurement and analysis tools that can handle and extract groups of events effectively and isolate the researcher from the now onerous task of database management and metadata collection necessary for validation and error analysis. Lack of information-management robustness or loss of metadata can lead to incorrect calibration results in addition to increasing the data-management burden. To address these issues, we have succeeded in automating several aspects of collection, parsing, reconciliation, and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. Future software automation tasks will continue to leverage database and information-management technologies in addressing additional scientific calibration research tasks.

OBJECTIVES

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next-generation data acquisition. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E program calibration efforts. The current task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration-product construction are susceptible to automation and future economies. Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. We have succeeded in automating many of the collection, parsing, reconciliation, and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. In order to fully exploit voluminous real-time data sources and support new requirements for time critical modeling, simulation, and analysis, a more scalable and extensible computational framework will be required.

RESEARCH ACCOMPLISHED

The primary objective of the Scientific Automation Software Framework (SASF) efforts is to facilitate development of information products for the Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) regionalization program. The SASF provides efficient access to, and organization of, large volumes of raw and derived parameters, while also providing the framework to store, organize, integrate and disseminate information products for delivery into the National Nuclear Security Administration Knowledge Base (NNSA KB). The current framework supports integration, synthesis, and validation of the various different information types and formats required by each of the seismic calibration technologies (Figure 1). For example, the seismic location technology requires parameter data (site locations, bulletins), and time-series data (waveforms) and produces parameter measurements in the form of arrivals, gridded geospatially registered corrections surfaces and uncertainty surfaces through the use of various tools and information-processing frameworks (relational databases (RDBs), Geographical Information Systems (GISs), and associated product and data visualization and data management tools (e.g., RBAP, KBALAP, KBCIT, DM).

These information-management and scientific automation tools are used together within specific seismic calibration processes to support production of tuning parameters for the United States Atomic Energy Detection System operated by the Air Force (Figure 2). The calibration processes themselves appear linear (Figure 2), beginning with data acquisition and extending from reconciliation, integration, and measurement and simulation to construction of calibration / run-time parameter products. Efficient production of calibration products, however, requires extensive synergy and synthesis not only between data-types (Figure 1), but also between measurements and results derived from the different calibration technologies (e.g., Location, Identification, Detection) (Figures 1 and 2). Even with successful implementation of automation within many of the individual steps, the current infrastructure will not scale to handle order-of-magnitude additional data or extend to handle time-critical data acquisition or analysis. This lack of scalability and flexibility limits efficient production and delivery of run-time calibrations to the operational seismic monitoring pipeline (Figure 2, bottom) as a large manual effort is still required to acquire and integrate streaming (10–20 GB/day) signals with associated metadata. This synergy and synthesis between complex tools and very large datasets is critically dependent on having a scalable and extensible unifying framework. These requirements of handling large datasets in diverse formats and facilitating interaction and data exchange between tools supporting different calibration technologies led to an extensive scientific automation software engineering effort to develop an object-oriented database-centric framework (Figure 3) as a unifying foundation.

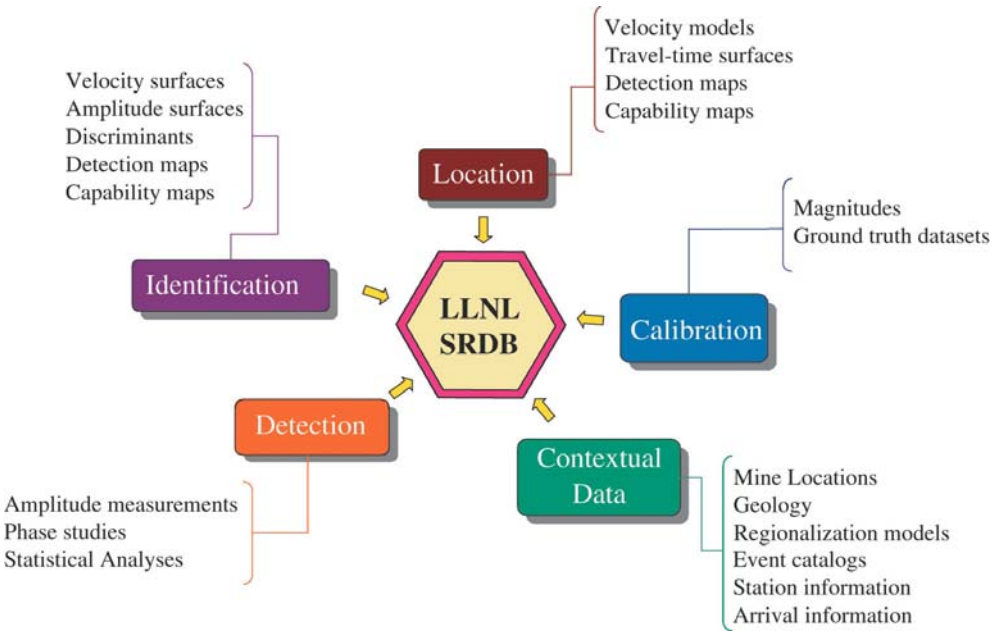


Figure 1. The Scientific Automation Software Framework provides a unifying framework for contextual/reference data and information products.

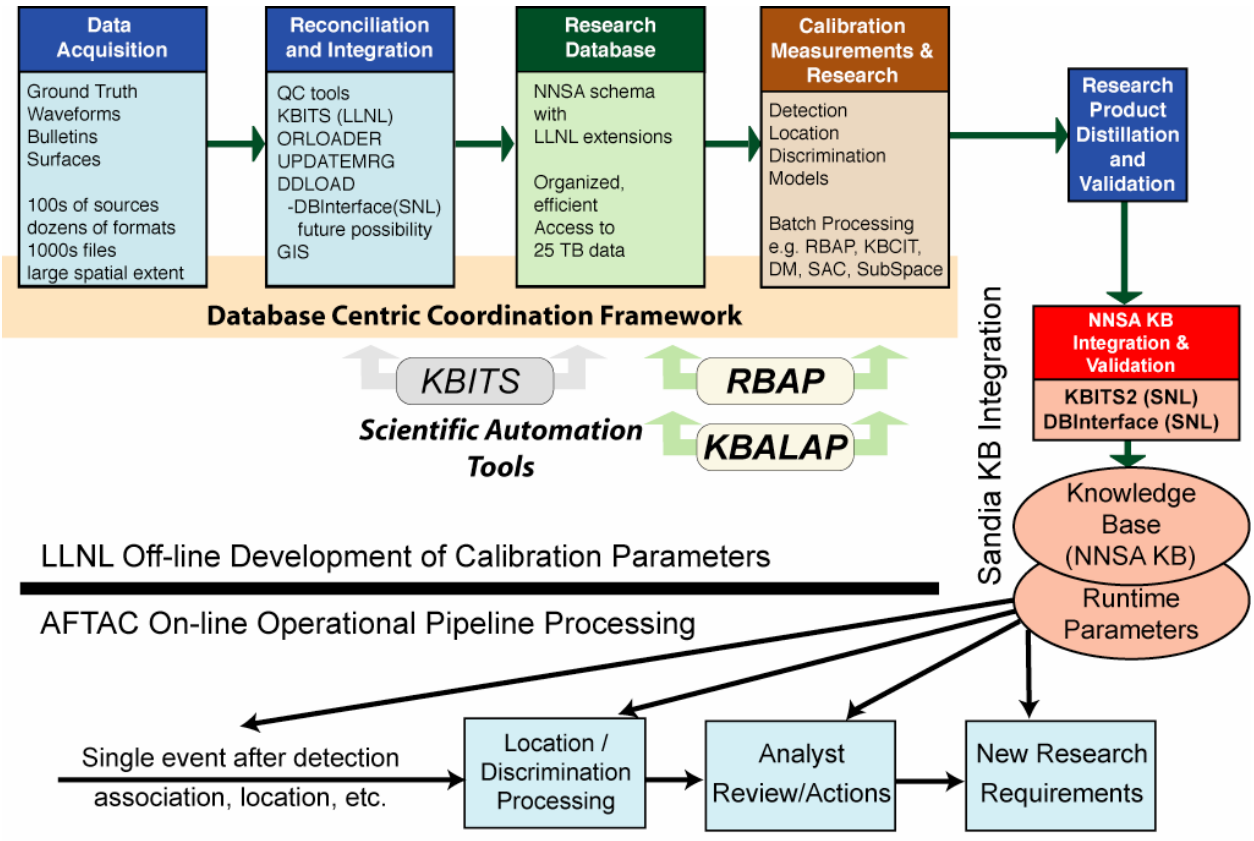


Figure 2: Summary of the processes of data collection, research, and integration within the LLNL calibration process that result in contributions to the NNSA KB. The relationships of the current LLNL calibration tools, scientific automation tools, and database coordination framework to those involved in the assembly of the NNSA KB are delineated.

Scientific Automation Software Tools

Information products created using the Lawrence Livermore National Laboratory (LLNL) Seismic Research Database (SRDB) may be grouped under two major categories or tiers: Tier 1 - primary data products, and Tier 2, derived products. In order to calibrate seismic monitoring stations, the LLNL SRDB must incorporate and organize the following categories of primary and derived measurements, data, and metadata:

Tier 1: Contextual and Raw Data
Station Parameters and Instrument Responses
Global and Regional Earthquake Catalogs
Selected Calibration Events
Event Waveform Data
Geologic/Geophysical Datasets
Geophysical Background Model

Tier 2: Measurements and Research Results
Phase Picks
Travel-time and Velocity Models
Rayleigh and Love Surface Wave Group Velocity Measurements
Phase Amplitude Measurements and Magnitude Calibrations
Detection and Discrimination Parameters

Automating Tier 1

Corrections and parameters distilled from the calibration database provide needed contributions to the NNSA KB for the Middle East/North Africa/Western Europe (ME/NA/WE) region and will improve capabilities for underground nuclear explosion monitoring. The contributions support critical functions in detection, location, feature extraction, discrimination, and analyst review. Figure 2 outlines the processes of data collection, research, and integration within the LLNL calibration process that result in contributions to the NNSA KB and the relationship of the LLNL calibration tools to those involved in the assembly of the NNSA KB or within the AFTAC operational pipeline. Within the major process steps (data acquisition, reconciliation/integration, calibration research, product distillation) are many labor intensive and complex steps. The previous bottleneck in the calibration process was in the reconciliation/integration step (Figure 2). This bottleneck became acute in 1998, and the KBITS suite of automated parsing, reconciliation, and integration tools for both waveforms and bulletins (ORLOADER, DDLOAD, UpdateMrg) was developed. The KBITS suite provided the additional capability required to integrate data from many data sources and external collaborations. Data volumes grew from the 11,400 events / 1 million waveforms in 1998 to the 6 million events / 70 million segmented waveforms and terabytes of continuous data today (e.g., Ruppert et al., 1999, Ruppert et al., 2004). This rapid increase in stored parameters soon led to new bottlenecks hindering rapid development and delivery of calibration research.

Automating Tier 2

As the number of data sources required for calibration increased in number and source location, it became clear that the manual and labor-intensive process of humans transferring thousands of files and unmanageable metadata could not keep the KBITS software fed with data to integrate, nor could the seismic research efficiently find, retrieve, validate, or analyze the raw parameters necessary to effectively produce seismic calibrations in an efficient manner. Significant software engineering and development efforts were applied to address this critical need to produce software aids for the seismic researcher. Two scientific automation tool prototypes (RBAP, KBALAP) (Figure 2) are under development for seismic location and seismic identification calibration tasks.

Both of these prototypes include methods and aids for efficiently extracting groups of events and waveforms from the millions contained in the SRDB and making large numbers of measurements with metadata in a batch mode. The concept of event sets (groups of related seismic events or parameters that can be processed together, e.g., either station centric or event centric) was introduced as previous SAC scripts and macros could not scale to the task.

The KBALAP Program

The Knowledge Base Automated Location Assessment and Prioritization (KBALAP) program is a set of database services and a client application that combine to efficiently produce location ground truth (GT) data that can be used in the production of travel time correction surfaces and as part of the preferred event parameters used by other tools in our processing framework.

The part of KBALAP that runs as a database service is responsible for evaluating bulletin and pick information as it enters the system to identify origin solutions that meet predefined GT criteria with no further processing, and to identify events that would likely meet a predefined GT level if a new origin solution were produced using available arrivals. The database service is also responsible for identifying events that should have a high priority for picking based on their existing arrival distribution and the availability of waveform data for stations at critical azimuths and distances.

The interactive portion of KBALAP has three principal functions. These are

- interactive production of GT origins through prioritized picking and location,
- interactive specification of GT-levels for epicenter, depth, origin time, etype, and
- batch-mode location of externally-produced GT information.

The first of these capabilities allows the user to view epicenters and GT information on a map based on selection criteria input by the user. The user can select any GT or potential GT event and observe the distribution of stations with picks and stations with available waveforms. The user can select any station with available waveforms and open a picker with any current picks displayed. There the user can adjust existing picks, add new picks, mark bulletin picks as unusable, and relocate the event. A new GT level is calculated, and the user can choose to accept that origin solution and GT level or continue working with other stations.

The interactive GT entry mode of KBALAP allows the user to retrieve information about a specific event and add or update that event's GT parameters. The program can also create a new event with a GT level for cases where epicenter, time, depth and magnitude GT data are available. Similarly, the batch mode part of the program allows specification of flat files containing GT data for events already in the database.

The RBAP Program

The Regional Body-wave Amplitude Processor (RBAP) is a software tool to help automate the process of making amplitude measurements of regional seismic phases for the purpose of calibrating seismic discriminants at each station. RBAP generates station-centric raw and Magnitude Distance Amplitude Correction (MDAC) corrected Pn, Pg, Sn and Lg amplitudes along with their associated calibration parameters (e.g., phase windows, MDAC values, reference events, etc.) in database tables. It strictly follows the Working Group (WG) 2 standardized processing described in the MDAC White Paper (Walter et al., 2003) and it replaces the original collection of scripts described by Rodgers (2003). RBAP has a number of advantages over the previous scripts. It is much faster, significantly easier to use, scales more easily to a larger number of events and permits efficient project revision and updating through the database.

RBAP integrates the functions of the modules in the previous LLNL scripts into a single program that is designed to perform the amplitude measurement task efficiently and to require a minimum effort from the users for managing their data and measurements. For well-located events with pre-existing analyst phase picks, the user reviews for quality control and then generates all the amplitudes with just a few mouse clicks. For events needing more attention, the user has complete control over the process (e.g., window control, ability to mark bad data, define regions, define MDAC parameters, and define the events to be used in the overall calibration process). RBAP shortens the time needed by the researcher to calibrate each station while simultaneously allowing an increase the number of events that can be efficiently included. RBAP is fully integrated with the LLNL research database. Data are always read directly from the appropriate tables in the research database rather than from a snapshot, as was done in the previous system. All RBAP result tables have integrity constraints on the columns with dependencies on data in the LLNL research database. This design makes it very difficult for results produced by RBAP to be stale and also ensures that as the research database expands, RBAP automatically becomes aware of new data that should be processed. RBAP initial users will be LLNL WG 2 members working on Integrated Research Products for FY04.

Some RBAP Key Features

• *Based on WG 2 Standardized Algorithm*
-RBAP is built on the WG 2 standardized body-wave amplitude measurement algorithms documented in the "MDAC White Paper" (Walter et al., 2003). Its results are completely consistent with the last version of the LLNL scripts (Rodgers, 2003) that were vetting in the February 2003 WG 2 exercise between LLNL, LANL, and AFTAC.

• *Fast and Efficient Calibration*
RBAP is self-contained and optimized for station-centric body-wave processing. "Good" events can be handled with just a few mouse clicks. The researcher has direct control over key calibration parameters within the tool such as phase amplitude windows and migration, marking bad segments, defining distinct geophysical regions, event types to process, etc. We expect RBAP to provide roughly a factor of 5 increase in calibration speed compared with the original scripts, enabling us to calibrate more stations, with more events per station.

• *Project Management*
RBAP is designed so that a calibration project can be put down for a day, month, or year, and easily picked up, by the same researcher or a new one. All processing metadata are saved and events are easily tracked as processed, unprocessed, or outside the current project definitions. This allows a researcher to efficiently work through a huge data list without repetition and to easily identify and incorporate new events as they become available in the database.

• *Utilizes Database for Up-to-Date results*
RBAP can draw on the latest calibration parameters being generated by other working groups, such as the most recent phase picks, relocations, magnitudes, instrument response information, or event type ground truth.

• *Batch Processing*
RBAP is designed to allow simple batch updating of the amplitude results, whether the change is small (e.g., one-event is relocated) or large (instrument response is changed, affecting all events).

Database Centric Coordination Framework

As part of our effort to improve our efficiency, we have allowed researchers to easily share their results with one another. For example, as the location group produces GT information, that information should become available for other researchers to use. Similarly, phase arrival picks made by any qualified user should also become immediately available for others to use. This concept extends to the sharing of information about data quality. It should not be necessary for multiple researchers to have to repeatedly reject the same bad data. Rather, once data are rejected because of quality reasons, they should automatically be excluded from processing by all tools. We are implementing this system behavior using database tables, triggers, stored procedures, and application logic. Although we are at the beginning of this implementation, we have made significant progress over the last year with several kinds of information sharing using the new Database Centric Coordination Framework. These are discussed below.

Significant software engineering and development efforts have been applied successfully to construct an object-oriented database framework that provides database-centric coordination between scientific tools, users, and data (Figure 3). A core capability this new framework provides is information exchange and management between different specific calibration technologies and their associated automation tools such as Seismic Location (e.g. KBALAP), seismic identification (e.g., RBAP), and data acquisition / validation (e.g., KBITS). A relational database (ORACLE) provides the current framework for organizing parameters key to the calibration process from both Tier 1 (raw parameters such as waveforms, station metadata, bulletins, etc.) and Tier 2 (derived measurements such as ground-truth, amplitude measurements, calibration and uncertainty surfaces etc). Efforts are underway to augment the current relational database structure with semantic graph theory structured queries for handling complex queries.
Seismic calibration technologies (location, identification, etc.) are connected to parameters stored in the relational database by an extensive object-oriented multi-technology software framework (Figure 3, middle) that include elements of schema design, stored procedures, real-time transactional database triggers, and constraints, as well as

coupled Java and C++ software libraries to handle the information interchange and validation requirements. This software framework provides the foundation upon which current and future seismic calibration tools may be based.

Sharing of Derived Event Parameters

We have long recognized the inadequacies of the CSS3.0 origin table to serve as a source of information about the "best" parameters for an event. One origin solution may have the best epicenter but poor information on other parameters.

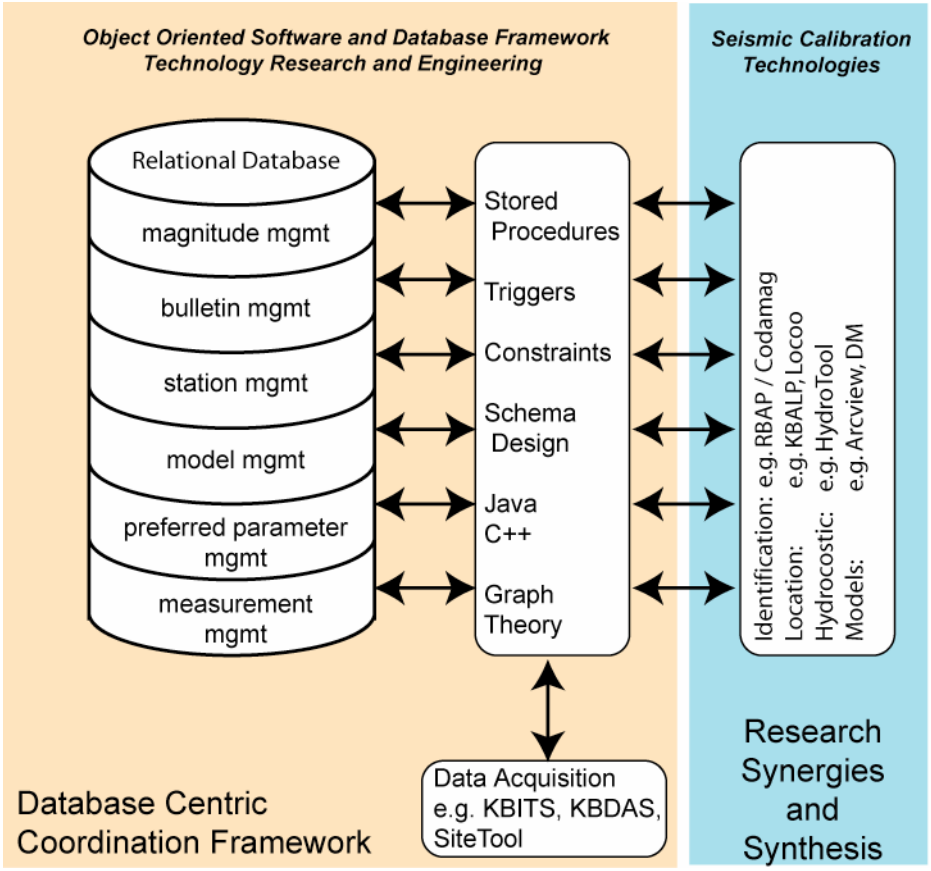


Figure 3. Overview of the Database Centric Coordination Framework that provides the enabling information technology to allow synergy and synthesis of data and calibration technologies for the efficient production of calibration deliverables.

Another may have the correct event type but be poor in other respects, and so on. We had discussed producing origin table entries, with our organization as the author, but that approach has difficulties. Different groups would have responsibility for different fields in the origin. Because their information would not be produced in synchronization, we would always have to be either updating the preferred origin or producing new preferred origins. Also, there would be difficulties in tracking the metadata associated with each field of the preferred origin. Our solution was to create a set of new tables and associated stored procedures and triggers that collectively maintain the "best" information about events.

Enhancements to Efficiency through Cluster-Based Computing

We have begun to leverage scalable and reconfigurable cluster computing resources to improve the efficiency of our computational infrastructure. Just as the database-centric approach to information management provided important gains in efficiency, we needed to move to a different computational paradigm to provide the computational power

necessary during calibration production and research. We have begun developing a set of flexible and extensible tools that are platform independent and parallelizable. These research tools will provide an efficient data processing environment for all stages of the calibration work flow, from data acquisition through making measurements to calibration surface preparation. This scalable and extensible approach (Figure 4) will result in more coupled and dynamic work flow in contrast to the linear work flow of the past, and allow more interaction between data, model creation, and validation processes.

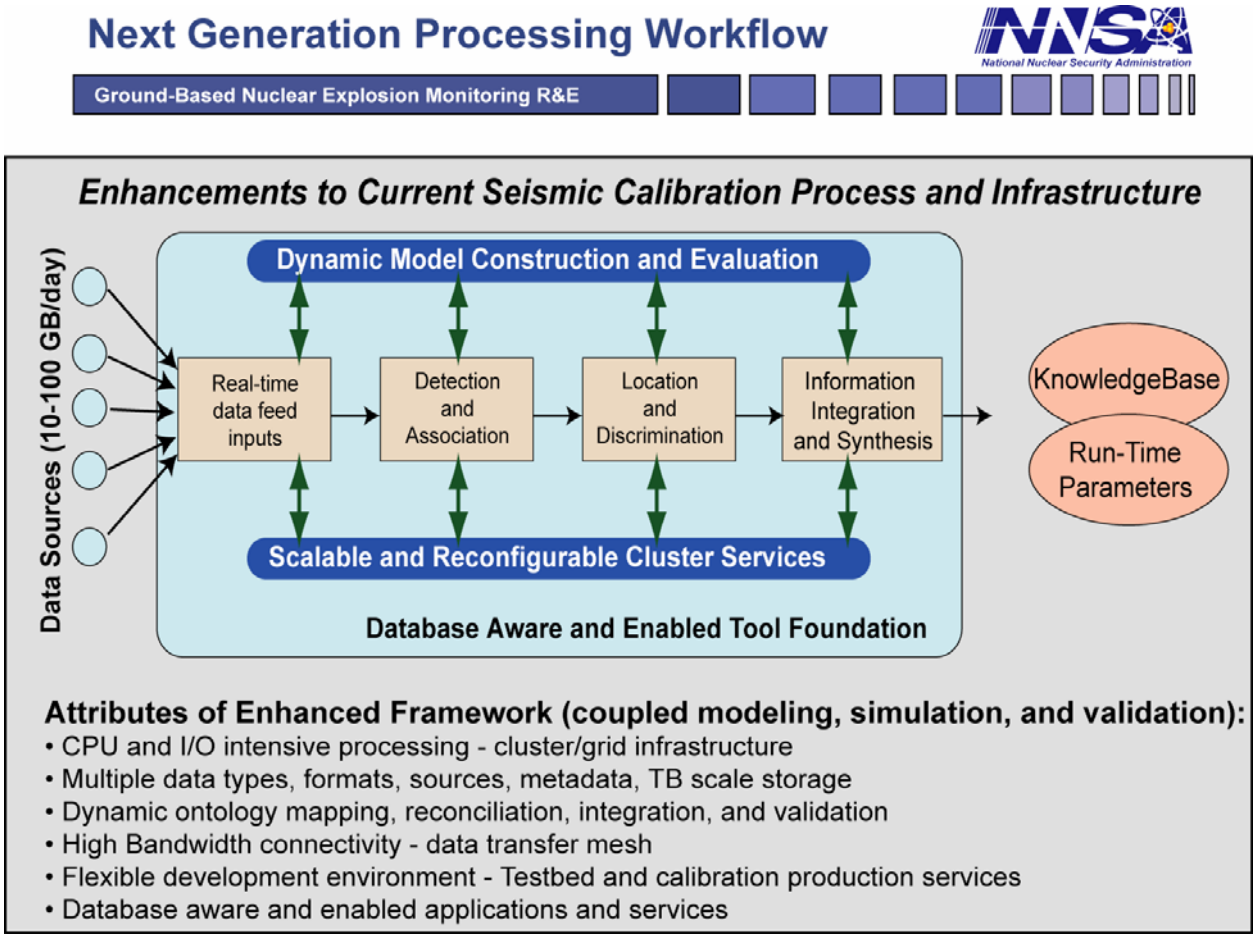


Figure 4. Overview of the cluster based calibration workflow under design and development.

Initial development and modification of existing codes and algorithms of the cluster based computing environment, has yielded significant efficiency improvements in RBAP and other measurement tools. Modification of RBAP to incorporate threads to isolate computationally intensive operations has provided a more interactive and responsive environment for the researcher and has laid the ground work for moving the threads to cluster-based computing resources. Other areas under investigation for taking advantage of cluster resources are for waveform correlation and subspace detector work, in addition to providing the ability to efficiently perform large-scale event relocations to evaluate ground truth and model calibrations.

CONCLUSIONS AND RECOMMENDATIONS

We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next-generation data acquisition. By combining both a database centric information management system coupled with scalable and extensible cluster-based computing, we have begun to leverage a high-performance computational framework to provide increased calibration capability. These new software and scientific automation initiatives

could directly support our current mission, including rapid collection of raw and contextual seismic data used in research, providing efficient interfaces for researchers to measure/analyze data, and providing a framework for research dataset integration. The initiatives would improve time-critical data assimilation and coupled modeling/simulation capabilities necessary to efficiently complete seismic calibration tasks. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E program calibration efforts.

ACKNOWLEDGEMENTS

We acknowledge the assistance of the LLNL computer support unit in implementing and managing our computational infrastructure. We thank Jennifer Aquilino and Laura Long for their assistance in configuration and installation of our Linux cluster.

REFERENCES

Ruppert, S., T. Hauk, J. O'Boyle, D. Dodge, and M. Moore (1999), Lawrence Livermore National Laboratory's Middle East and North Africa Research Database, in *Proceedings of the 21st Seismic Research Symposium: Technologies for Monitoring The Comprehensive Nuclear-Test-Ban Treaty*, LA-UR-99-4700, Vol. 1, pp. 234-242, Las Vegas, NV.

Ruppert, D. Dodge, A. Elliott, M. Ganzberger, T. Hauk, E. Matzel, and F. Ryall (2004), Enhancing seismic calibration research through software automation, in *Proceedings of the 26th Seismic Research Review: Trends in Nuclear Explosion Monitoring*, LA-UR-04-5801,Vol. 2, pp. 780-789, Orlando, FL.

DATA AND TOOLS TO SUPPORT NUCLEAR EXPLOSION MONITORING RESEARCH AND DEVELOPMENT

Robert Woodward, Michael Skov, Manochehr Bahavar, Geoff Davis, and Yu-Long Kung

Science Applications International Corporation

Sponsored by Army Space and Missile Defense Command

Contract No. DASG60-03-C-0009

ABSTRACT

The nuclear explosion monitoring research and development (R&D) community can access raw data and unique research databases using state-of-the-art data browsing and selection tools as part of the Research and Development Support Services (RDSS) project operated by the United States (US) Army Space and Missile Defense Command Monitoring Research Program.

The waveform data archive maintained by the RDSS provides direct and immediate access to 100% of the data in the archive. A high-speed spinning disk mass-storage system is used to store all the waveforms in the archive (roughly 12 Tb of seismic, hydroacoustic and infrasound waveforms from over 300 locations worldwide). Every waveform in the archive can be directly and instantly accessed, making it possible to provide data users with a range of new tools and services that go beyond traditional data center functions. User interaction with the data archive is provided through several mechanisms, including a web-based, "e-commerce-like," interface. This web-based interface allows users to visually browse data (waveforms) or data products (e.g., bulletins), select data or data products to be placed in a "shopping cart," and then provides the capability to manage and download the selected objects. A waveform viewer runs under a standard web browser and provides the capability to display multiple traces, perform filtering, zooming, and scrolling. Any waveform in the entire data archive can be viewed using this tool. The user can select waveforms from the display for subsequent download. A download manager then provides the ability to download the various products a user has selected during a session (or sessions—the download manager can manage product selections made during multiple sessions). Any data in the archive can be downloaded in either CSS3.0 or SAC formats.

The waveform data archive also supports data intensive computing by providing direct access to all data files in the archive via RDSS servers. Users who load software on these servers can perform experiments on any desired cross-section of the waveform archive, with no need to "stage" data. Further, it is not necessary to make special modifications to scientific software, since the data are stored in the archive in a common, analysis ready format. An example of a recent large-scale, data intensive experiment was the analysis of ambient noise at infrasound stations. This experiment required computing noise values at multiple infrasound stations, at multiple times per day, for every day of a 2-year period. Over a million discrete (noncontiguous) time series were analyzed for this study.

The RDSS also provides a range of value-added R&D databases, which are a significant resource for the US R&D community. These databases are accessible through interactive, web-based tools and bring together a wide range of open-source data into individual, well-organized packages. For example, the recently expanded infrasound database draws on a unique collection of waveforms (many of which are not archived anywhere else) from infrasound arrays operated by the Department of Energy, the International Monitoring System, and other organizations. These data go back to 1995, and include data from sites on every continent. The database includes acoustic recordings of a variety of natural and man-made events, as well as historical recordings of atmospheric nuclear explosions. The seismic R&D databases include the Nuclear Explosion Database, an exhaustive collection of metadata and waveforms from nuclear tests; the Ground Truth Database, containing a wide range of carefully selected and quality-controlled events for ground-truth (GT) levels of 0 to 15 km; and region-specific databases for the Lop Nor, China, region and the Arctic region.

In this paper we provide a summary of the major RDSS data assets and resources. We provide examples and descriptions of the types of data and metadata in each database, and provide information on how these resources can be accessed.

OBJECTIVE

The objective of the RDSS is to support the nuclear explosion monitoring research and development community with a wide range of data, state-of-the-art data access tools, and value-added datasets.

RESEARCH ACCOMPLISHED

During the current contract we have pursued initiatives in three primary areas with the goal of extending existing resources or developing new resources which will provide direct benefit to the nuclear explosion monitoring R&D community. First, we have expanded and improved the waveform archive, concentrating especially on creating a high-speed archive and new access tools which fully exploit the capabilities of the new system. Second, we have improved our existing, value-added R&D databases and have created several new databases. Third, we have developed new tools, using an application service provider model, to provide remote users with access to sophisticated R&D software. In the following sections we describe these developments in greater detail.

Waveform Archive and Data Access

The RDSS maintains a large archive of waveform data from seismic, hydroacoustic, and infrasound stations. Data in the archive go back to 1995 (earlier for some stations) for stations distributed worldwide (Figure 1).

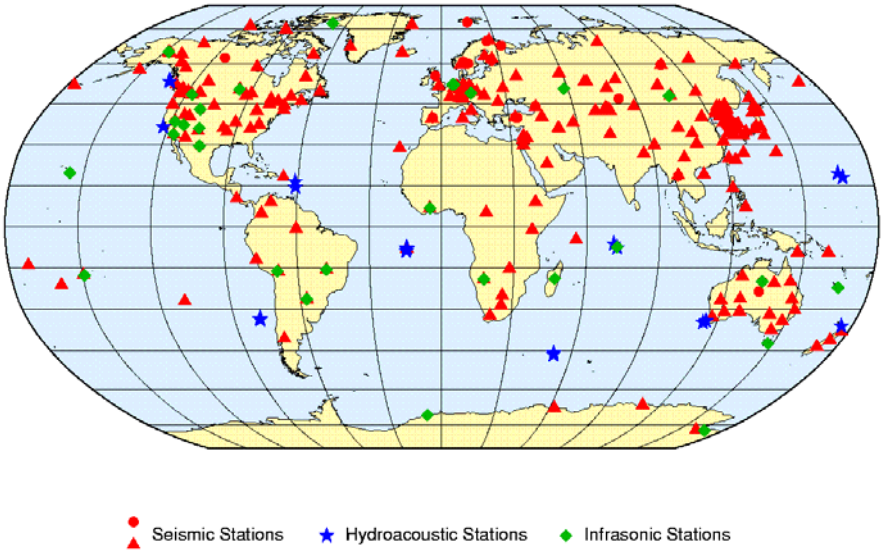


Figure 1. Seismic arrays (circles) and three-component stations (triangles) and hydroacoustic and infrasonic arrays (stars and diamonds, respectively) for which waveform data are available from the RDSS archive. For many of these stations and arrays the data are continuous over periods of several years.

The waveform archive now exceeds 14 Tb of data. The entire data archive is hosted on a spinning-disk mass storage system (RAID disk farm). This system provides access to the entire archive that is at least two orders of magnitude faster than its predecessors and makes very large scale experiments and data visualizations possible.

The RDSS website is continuously enhanced to improve data access for the research community. During the past year we have initiated work to integrate the suite of tools and interfaces available on the web site into an "e-commerce-like" interface. The goal of this effort is to provide a unifying theme for all user interactions (i.e., data selection, data downloads, etc.) with the site that is based on concepts familiar from e-commerce (on-line shopping) websites. The goal of this web model is to allow users to visually browse data (waveforms) or data products (e.g., bulletins), select data or data products to be placed in a "shopping cart," and then provide the capability to manage and download the selected objects.

We have substantially improved the web-based waveform viewing capability. Last year we introduced an online waveform viewing capability which allowed users to view any waveform in the entire data archive. This capability has been extended to provide mult-itrace support and enhanced waveform display control. The updated waveform

viewer (Figure 2) allows users to simultaneously view multiple channels of data from a single station or to view data channels from different stations. The waveform tool is not intended as an analysis tool, but rather as a waveform browser to streamline the process of locating, previewing, selecting, and acquiring waveforms of interest. However, basic scroll, zoom and filter functions are available. The waveform viewer is accessed through a standard web browser—no special installations or modifications on the user's (client's) machine or web browser are required.

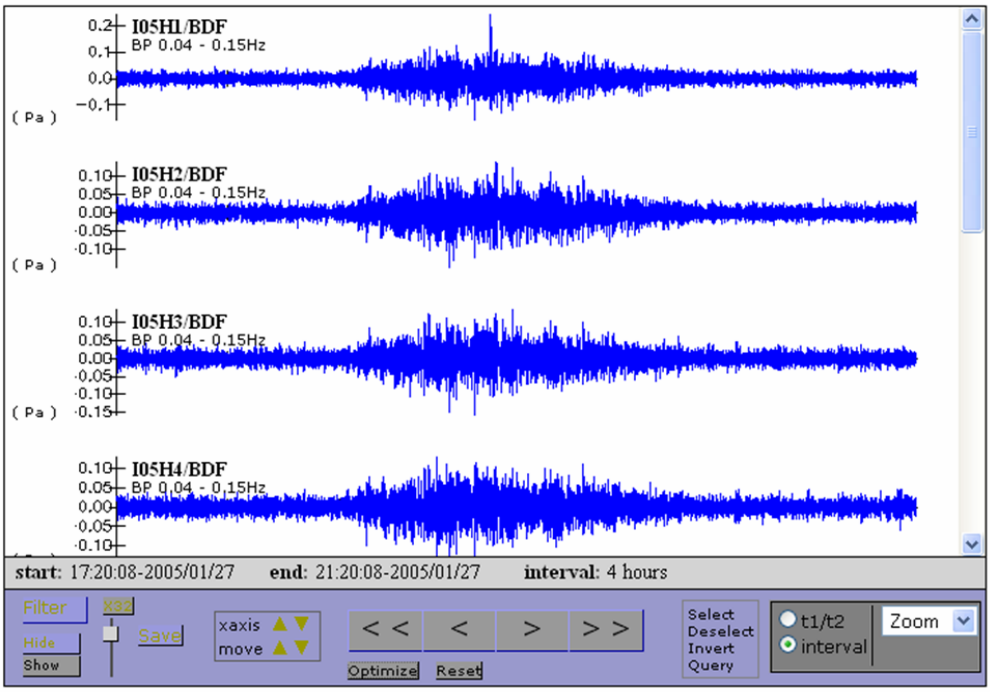


Figure 2. The web-based data viewing tool provides easy access to all data in the archive. Data on display are from station I05AU in Australia and show signals from the eruption of a volcano on Manam Island at approximately 38° distance. The data have been bandpass filtered between 0.04 and 0.15 Hz.

Once the desired data have been identified, the user "selects" these data (equivalent to placing items in a shopping cart on an e-commerce website). Data selections are stored in a private folder for later review, revision, and download (Figure 3). User preferences for data format and delivery method are recalled automatically, greatly reducing the effort required to obtain waveform data. For example, both CSS3.0 and SAC formats are available for waveform data. Multiple data selections may be packaged together into a single download and once the content of a particular data package is defined, the user may return to browsing while the data are collected and converted to the user's preferred format. A given data package may be downloaded more than once and prior downloads are recorded in a history folder, freeing researchers from the need to maintain permanent data warehouses on their computer.

An enhanced waveform data location service is in development as a replacement for the existing General Data Extractor (GDE) web service. Like the existing GDE, this tool will permit users to locate waveform data based on event information and other constraints such as group velocity limits and preferred station and channel selections. The replacement tool is enhanced to provide integration with both the new private folder infrastructure and the improved waveform browser—enabling the user to preview GDE-based waveform selections as well as manage selections via the user's private data folder.

A major goal in the development of these new tools is to support collaborative research by distributed research teams. Research teams might identify and define relevant data sets to be shared by members of the team. These data selections would then be available to any team member at any time without each individual requiring a personal copy of all data. Researchers would need only to download to their personal workstations those data subsets relevant to the current task. These downloaded data could be removed and replaced with other research team data as needed, freeing each team member from managing individual data sets.

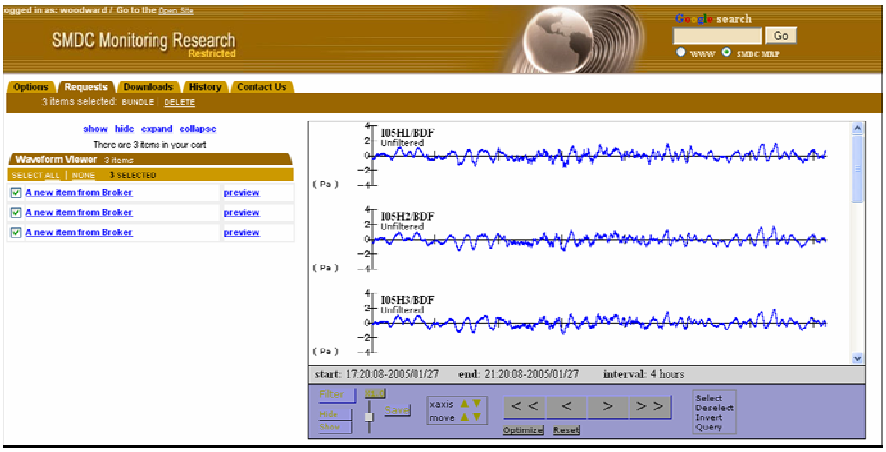


Figure 3. Using the data download manager, the user can review data in a "shopping cart." In this example, the data selected with the waveform viewing tool (e.g., the data shown in Figure 2) are being reviewed prior to downloading (note that in this view the data are shown unfiltered).

In addition to the new web-based data access tools described above, we provide backward compatibility for existing tools, such as the e-mail based *AutoDRM* software. In many cases the performance of these existing tools is greatly improved because of the quick response times of the spinning-disk mass-storage system.

Data-Intensive Computing

It is possible to perform massive, data-intensive experiments directly on the data in the archive. This is possible since the archived data are written on the mass-storage system in an analysis ready format (CSS3.0). The data can be read directly by application software—making it possible to perform experiments on any desired cross section of data in the archive, with no need to "stage" data in advance.

Recent studies of infrasound ambient noise (Bowman et al., 2005a, b) provide an excellent example of the types of data-intensive computations facilitated by the high-speed mass-storage system. To characterize ambient infrasound noise, power spectral density was measured for 28 of the 34 infrasound stations shown in Figure 4. Data were analyzed from January 20, 2003, through December 31, 2004, from 21 consecutive 3-minute segments of 20 sample per second data taken four times daily, beginning at 06:00, 12:00, 18:00, and 24:00 local time. This required computing power spectra for 1,476,309 data segments (Figure 5), comprising approximately 5.3 billion data samples. In a companion study (Bowman, 2005a, b), average meteorological conditions (wind speed, wind direction, temperature, and absolute pressure) were calculated for the same 3-minute windows at 21 of the IMS stations. This computation utilized approximately three million data segments.

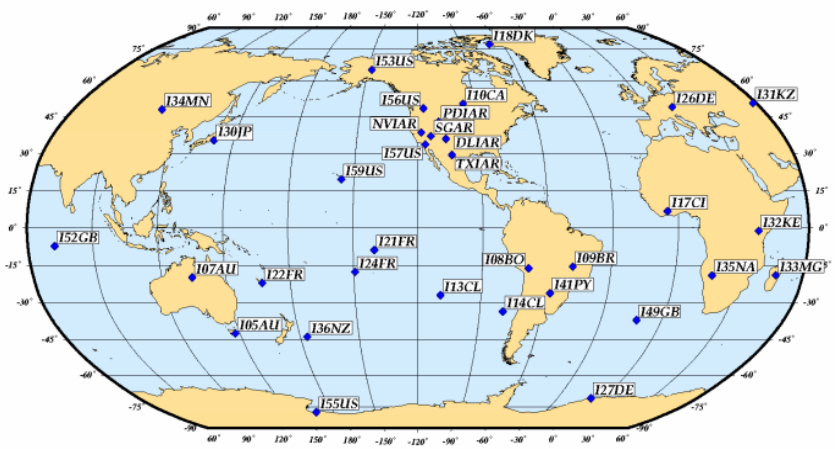


Figure 4. Data from 34 current infrasound stations are in the RDSS data archive.

The RDSS has a special server configuration to provide data intensive computing capability to the nuclear explosion monitoring R&D community. Remote users access a server which has direct (read-only) file access to the entire spinning-disk mass-storage system. Users essentially upload their code to the server rather than download data. The UNIX server is equipped with compilers, database access, and local storage for writing the results of computations. The user's code runs on the server and can read any waveform data in the archive (all data files are in CSS3.0 format). Researchers interested in utilizing this service should contact the RDSS.

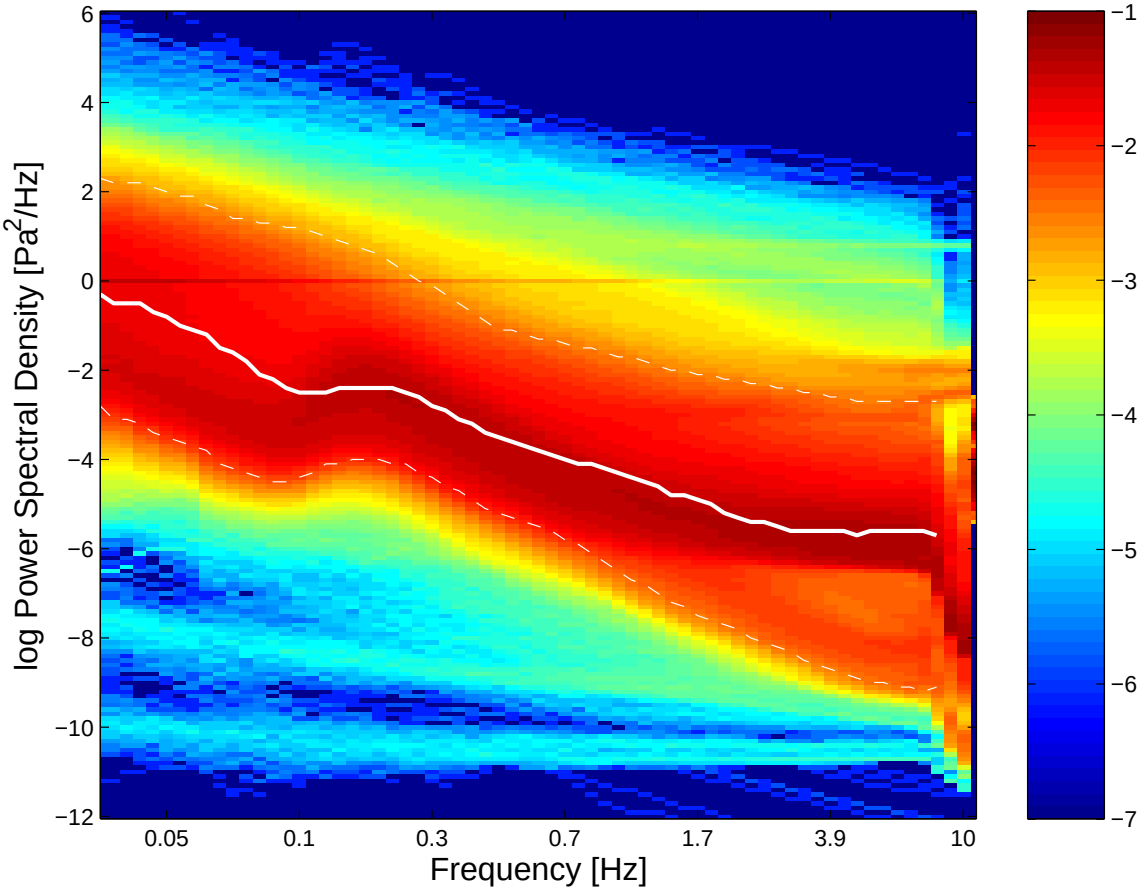


Figure 5. Probability density function of Power Spectral Density (PSD) for all time intervals at 25 infrasound stations. About 1.4 million spectra are binned in intervals of 0.1 log PSD. The color scale is proportional to the log of the number of spectra that fall within each bin. The solid white line shows the all-time network median for 15 stations having data for 1 year or longer, and the lower and upper dashed white lines show the low- and high-noise models, respectively (figure from Bowman et al., 2005b).

Value-Added Databases

The RDSS provides value-added R&D databases which are a significant resource for the US R&D community (Table 1). The data (waveforms and/or arrivals) included in these databases are extracted from the general RDSS waveform archive (described above) as well as from a wide variety of other open sources. The careful compilation of data, along with relevant metadata, is typically a challenging and time-consuming task. However, the resultant data sets provide unique, value-added resources which can support studies in nuclear explosion detection, location, and identification. We describe several of the value-added databases below. Users are encouraged to visit the Monitoring Research Program's RDSS web pages (http://www.rdss.info) for more details on all of the value-added databases described in Table 1.

Table 1. Summary of value-added research databases available through the RDSS.

Database	Description
Infrasound	Acoustic recordings from natural and man-made events, including atmospheric nuclear explosions. Data suitable for detection and source classification studies.
Nuclear Explosion	A vast compilation of seismic and infrasound recordings and arrival data from underground, underwater, and atmospheric nuclear explosions.
Ground Truth	Seismic phase arrival data for GT0 through GT15 events.
Lop Nor	Seismic waveforms (~100 GB) and phase arrival data for events and stations in the Lop Nor, China region. Includes recordings of larger nuclear explosions which have been scaled down to smaller yields and embedded in background noise. Provides multiple source location estimates obtained from different open sources. See Kohl et al. (2002).
Arctic Region	Seismic waveforms and phase arrival data for events and stations in the Arctic region. Provides multiple source location estimates obtained from different open sources.
Network Data Set	Database of event and waveform data for seismic and infrasound stations to support evaluation of detection and network processing performance.

The infrasound database has been significantly expanded during the past year. The database contains waveforms from infrasound arrays operated by the Department of Energy, the International Monitoring System, and other organizations (many of which are not archived anywhere else). The 34 infrasound arrays for which continuous data are currently in the database are shown in Figure 4. These data holdings go back to 1995 and include data from sites on every continent. The database also includes a research component, which contains recent and historic acoustic recordings of a variety of natural and man-made events, including atmospheric nuclear explosions (Table 2 and Figure 6). The signals in the infrasound database are suitable for a variety of signal detection and source characterization studies, such as the examination of such issues as seasonal propagation variability.

Table 2. Summary of events in the Infrasound Research Database. Numbers in black indicate current database holdings. Numbers in red represent data recently added, and numbers in blue indicate events to be added (i.e., in the queue).

Event Source Type	Events	Events with Arrivals	Arrivals	Station Waveforms
Nuclear explosion	63 (31) (8)			250 (110)
Chemical explosion	14 (4) (2)	10 (1)	17 (2)	26 (6)
Gas pipe explosion	3 (1)	2	3	5 (2)
Mine explosion	5 (5) (13)	5 (5)	7 (7)	5 (5)
Rocket launch	7	7	10	10
Bolide	19 (7) (8)	10	18	43 (19)
Rocket reentry	4	4	7	7
Volcano	29 (3)	1	1	29
Earthquake	3 (15)	3	4	4
Aircraft (not sonic boom)	1	1	1	1
Auroral wave	48	–	–	48
Gravity wave	5	–	–	5
Microbarom	29	–	–	29
Mountain wave	49	–	–	49
Unknown	18	14	14	18
Total	297 (48) (49)	57 (6)	82 (9)	529 (142)

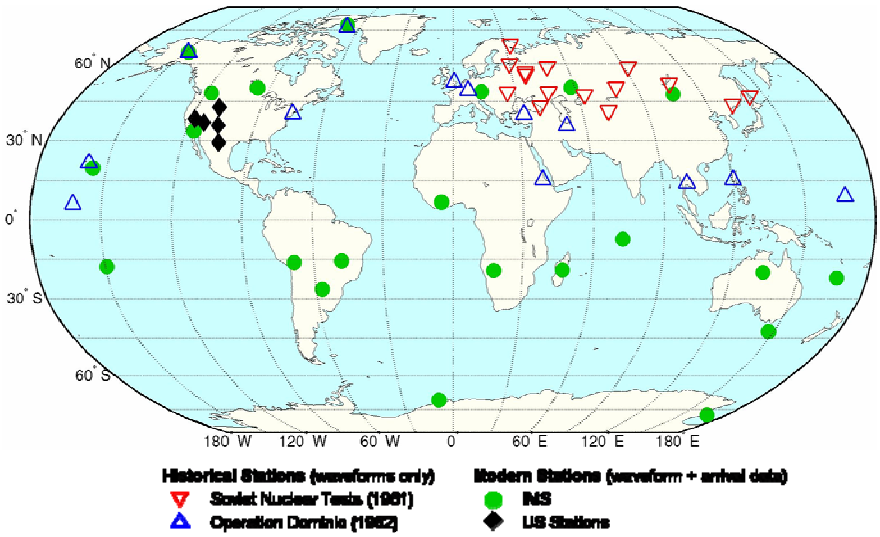


Figure 6. The Infrasound Research Database contains waveform data from a variety of historic and current station locations.

The Network Data Set (Kohl et al., 2005) provides a comprehensive database for evaluating detection and network processing performance. This "synthesized" network data set provides the means to evaluate signal processing algorithms on a large data set which has completely known characteristics. The data set utilizes seismic and infrasound signals from actual events. The signals are scaled to various sizes and embedded in clean background noise. Figure 7 shows the locations of the stations and events currently in the Network Data Set.

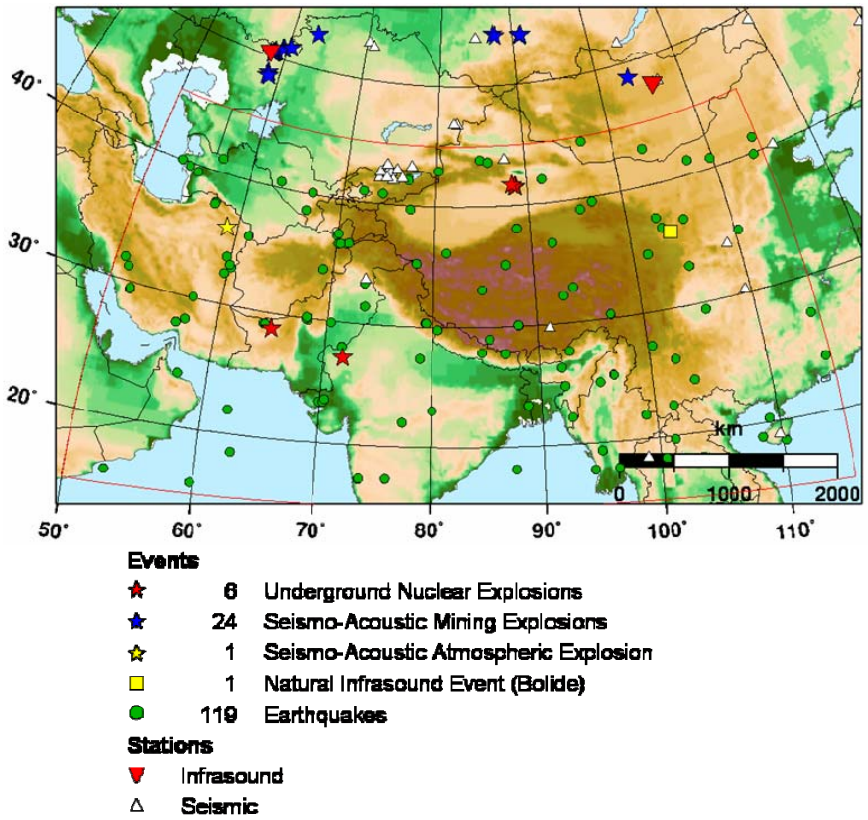


Figure 7. Locations of events and stations currently in the Network Data Set.

The Nuclear Explosion Database contains information on all reported nuclear explosions in the atmosphere, underground and underwater since the first nuclear explosive device was set off in 1945. The database gives the most accurate and complete unclassified information on time and place of both announced and presumed nuclear explosions and provides, whenever available, information on explosion yield, depth and shot medium (Table 3). In addition, the database contains a large archive of seismic (and some infrasound) recordings from about one-third of the explosions (Figure 8).

The explosion database was originally compiled in 1997 and has subsequently been maintained and updated. It currently holds data for 2041 explosions. The list of explosions in the data base was derived from more than 40 different data sources with information reported by both official, such as Department of Energy (DOE/NV-209, 2000), and non-official organizations and appearing in a variety of publications. A detailed description of the database has been compiled by Yang et al. (2000, 2003) and Bondár et al. (2001).

Table 3. Locations of stations for which nuclear explosion waveforms (seismic or infrasound) or arrival data are included in the Nuclear Explosion Database.

Explosion Environment	Total Number of Explosions	Number of Explosions with:				
		Shot Depth/ Height	Yield	Shot Medium	Arrival data	Waveforms
Underground	1516	632	1035	457	608	739
Underwater	10	7	9			1
Atmosphere	515	110	452		7	25

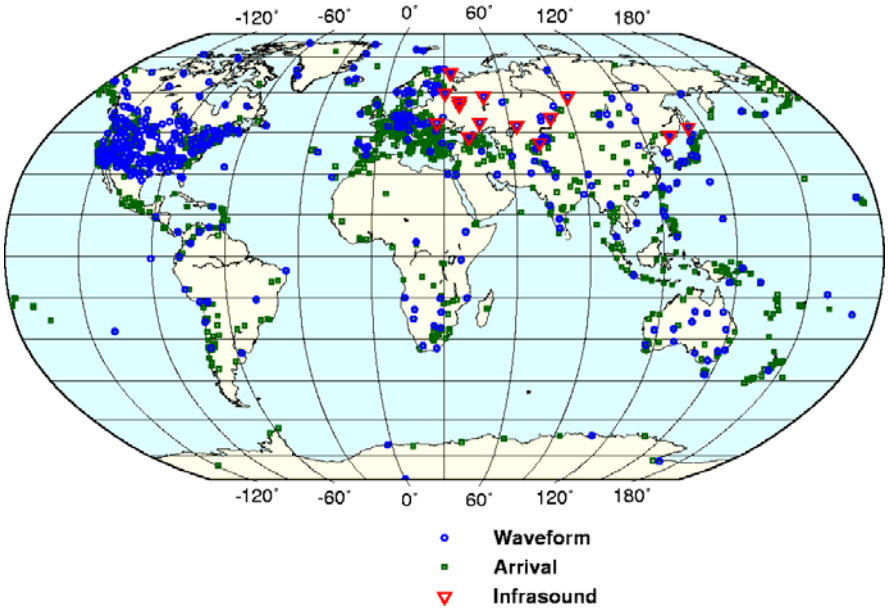


Figure 8. Locations of stations for which nuclear explosion waveforms (seismic or infrasound) or arrival data are included in the Nuclear Explosion Database.

The Ground Truth Database contains a collection of events of GT quality to support location calibration studies. The database includes nuclear and chemical explosions, mine blasts, rock bursts and earthquakes of global coverage. The earthquakes in the database all have shallow focus (less than 40-km focal depth), since deep earthquakes are of lesser interest from a nuclear monitoring point of view. Every event is accompanied with arrival data as well as references to document the source of information.

The events in the Ground Truth Database are distributed globally, though they are dominantly in the northern hemisphere (Figure 9). The database currently holds 13,379 GT0-15 events (Table 4) with some 900,000 associated phases (of which some 810,000 are defining) recorded at 3,771 stations. Residuals in the bulletins refer to IASP91 (Kennett and Engdahl, 1991) predicted travel times with the source fixed to the GT locations. Note that the categories GT7 and GT11 stand for events promoted to GT status after performing multiple event location on an event cluster. For details see Bondár et al. (2004), Engdahl and Bergman (2001), and Engdahl et al. (2002). The GT15 events (Pan et al., 2002) are included only to provide coverage on mid-oceanic ridge events in the mid-Atlantic ridge, the Carlsberg ridge and the Gulf of Aden.

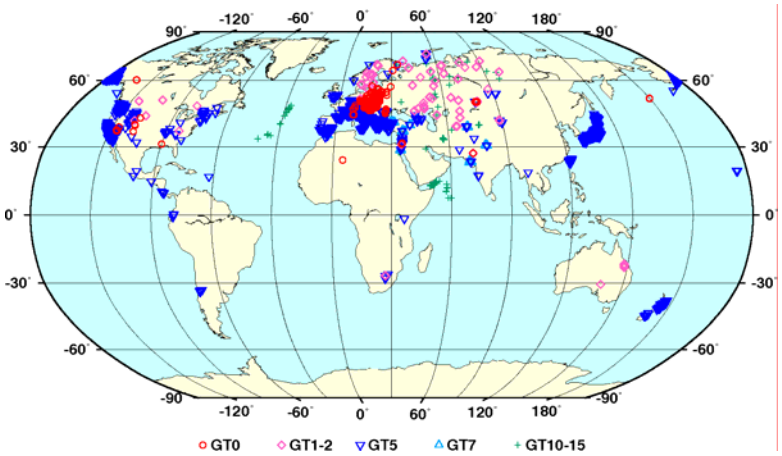


Figure 9. Locations of events in the Ground Truth Database.

Table 4. Summary of Ground Truth Database holdings, by source type and GT level.

Source	GT0	GT1-2	GT5	GT7	GT10	GT11	GT15	Total
Nuclear explosion	428	359	23	-	26	-	-	836
Chemical explosion	155	44	1	-	-	-	-	200
Mine blast, rock burst	-	141	58	-	-	-	-	199
Earthquake	-	-	11596	324	12	183	29	12144
Total	583	544	11678	324	38	183	29	13379

CONCLUSIONS AND RECOMMENDATIONS

We have described a variety of resources which are available to the US nuclear explosion monitoring research and development community.

A vast quantity of data is available to users via the spinning disk mass-storage system. New tools have been developed which make it possible to visually browse any waveform in the entire data archive. The waveform viewer runs under any standard web browser and provides the ability to view multiple channels of data from one or more stations. Data can be selected via the waveform viewer or with other tools. Selected data can then be downloaded via the data download manager, which provides user-specific configuration and download management.

The mass-storage system also supports data intensive computing experiments. Application software can access data directly from the mass-store—facilitating experiments which require large volumes of data from any combination of stations and time windows. Researchers interested in remotely accessing the data intensive computing infrastructure should contact the RDSS (e.g., via the website).

A wide range of value added databases have been produced. These databases represent unique assemblies of waveform data, arrival and event data, and metadata. The databases are all available on-line and provide valuable resources for use in detection, location, and identification studies.

We recommend that researchers in the nuclear explosion monitoring R&D community visit the MRP websites to learn more about the MRP's RDSS resources. The MRP maintains both an open website (http://www.rdss.info) and a restricted access website (at https://www.rdss.info; access limited to US government-authorized researchers).

ACKNOWLEDGEMENTS

We thank the members of the US Infrasound Team for making available high-quality data from the research stations and International Monitoring System stations which they operate.

REFERENCES

Bowman, J. R. (2005a), Meteorological Conditions at Infrasound Stations, *Inframatics,* 9.

Bowman, J. R. (2005b), Meteorological Conditions at Infrasound Stations, SAIC Technical Report SAIC-05/3000.

Bowman, J. R., G. E. Baker, and M. Bahavar (2005a), Ambient Infrasound Noise, *Geophys. Res. Lett.*, 32: L09803, doi:10.1029/2005GL022486.

Bowman, J. R., G. Shields, G. E. Baker, and M. Bahavar (2005b), Infrasound Station Ambient Noise Estimates and Models, SAIC Technical Report SAIC-05/3001.

Bondár, I., E. R. Engdahl, X. Yang, H. A. A. Ghalib, A. Hofstetter, V. Kirichenko, R. Wagner, I. Gupta, G. Ekström, E. Bergman, H. Israelsson, and K. McLaughlin (2004), Collection of a Reference Event Set for Regional and Teleseismic Location Calibration, *Bull. Seism. Soc. Am.* 94: 1528–1545.

Bondár, I., X. Yang, R.G. North, and C. Romney (2001), Location Calibration Data for CTBT Monitoring at the Prototype International Data Center, *Pure Appl. Geophys*. 158: 19–34.

DOE/NV-209 (2000), United States Nuclear Tests, July 1945 through September 1992, Rev. 15.

Engdahl, E. R. and E. A. Bergman (2001), Validation and Generation of Reference Events by Cluster Analysis, in *Proceedings of the 23rd Seismic Research Review: Worldwide Monitoring of Nuclear Explosions*, LA-UR-01-4454, Vol. 1, pp. 205–214.

Engdahl, E., E. Bergman, M. Ritzwoller, N. Shapiro, and A. Levshin (2002), A Reference Event Data Set for Validating 3-D Models, in *Proceedings of the 24th Seismic Research Review—Nuclear Explosion Monitoring: Innovation and Integration*, LA-UR-02-5048: 261–270.

Kennett, B. and E. R. Engdahl (1991), Travel Times for Global Earthquake Location and Phase Identification, *Geophys. J. Int*. 105: 429–465.

Kohl, B., R. North, J. R. Murphy, M. Fisk, and G. Beall (2002), Demonstration of Advanced Concepts For Nuclear Test Monitoring Applied to the Nuclear Test Site at Lop Nor, China, in *Proceedings of the 24th Seismic Research Review—Nuclear Explosion Monitoring: Innovation and Integration*, LA-UR-02-5048, Vol. 1, pp. 302–312.

Kohl, B., T. J. Bennett, I. Bondár, B. Barker, W. Nagy, C. Reasoner, H. Israelsson, P. Piraino (2005), Development of a Network Data Set for Evaluating Detection and Network Processing Performance, in current Proceedings.

Pan, J., M. Antolik, and A. Dziewonski (2002), Locations of Mid-Oceanic Earthquakes Constrained by Sea-Floor Bathymetry, *J. Geophys. Res.* 107: 2310, doi:10.1029/2001JB001588.

Yang, X., R. North, and C. Romney (2000), CMR Nuclear Explosion Database (Revision 3), CMR Technical Report CMR-00/16.

Yang, X., R. North, C. Romney, and P. Richards (2003), Worldwide Nuclear Explosions, in *International Handbook of Earthquake and Engineering Seismology*, Vol. 81B, W. H. Lee, H. Kanamori, P. Jennings, and C. Kisslinger, Eds., Academic Press.

A POTENTIAL NEW NUCLEAR EXPLOSION MONITORING SCHEMA

Christopher J. Young and Sanford Ballard

Sandia National Laboratories

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. DE-AC-04-94-AL85000

ABSTRACT

Relational database schemas are widely used in nuclear explosion monitoring applications (Carr, 2004) and they work well for many aspects of monitoring, but have some limitations that can only be addressed through fundamental redesign. In this paper we present a new seismic schema that captures all of the information in the current core schema, but is better normalized, more extensible, and provides a better basis for next generation seismic monitoring research and system development.

Past modifications of the core schema have typically been minor, avoiding significant changes to the existing tables because software that interfaces with them may have to be modified also. Introducing custom tables is equally problematic and similarly avoided. In the short term, keeping the original tables makes sense because software modifications can be expensive. For conducting new seismic monitoring research and developing next generation monitoring software systems, however, keeping the current table structure can become a significant impediment.

Using the Object Role Modeling (ORM) component of Visio Enterprise Architect Edition, and starting from the basic object fact relationships, we have developed a data model and resultant schema that supports many-to-many event-to-origin relationships, completely captures complex model prediction information, properly records array membership, and properly normalizes network magnitude storage, to name just a few features. By using the Visio tool, we are assured that our resultant schema is highly normalized, that all tables are linked with settable primary, foreign and unique keys, and that important constraints are enforced within the database via stored procedures. The resulting schema favors efficiency and maintainability, but ease of interaction with the information in the database can still be achieved through the use of views. We have made great efforts to make our design easily extensible without modifying the schema, but should modifications be required they can be made easily through the ORM model. Because we include the ORM as part of the schema design, there is no confusion in our schema about why a table or column is included: every feature of the schema can be related directly back to a simple fact statement that can be validated by any seismic monitoring domain expert. To verify the accuracy of our new schema, we have developed sql scripts to set up the old schema tables as views of the new tables.

Our initial development has focused on seismic monitoring, but we have made efforts to keep the schema as generic as possible so that it can be used to capture similar types of information for other ground-based monitoring technologies (e.g., hydroacoustic and infrasound). Should our proposed schema be chosen for use with real monitoring systems, large amounts of old-format data would need to be converted to the new schema. We show in this paper how this would be done with a simple custom utility.

OBJECTIVE

The basic table structure of the seismic schema currently used for nuclear explosion monitoring was developed more than 15 years ago (Anderson et al., 1990) and has changed very little since then (Carr, 2004). We have had many years of experience with this schema, both using it to hold data as well as designing increasingly sophisticated software to interface with it. For most basic purposes, the schema works well, but our work with developing the NNSA Knowledge Base has produced a list of problems that is both too long and of too fundamental a nature to address with simple fixes to the current tables. In particular, we think that the current design has become an impediment to modern software design practices (e.g., object oriented coding) that rely on a well-designed schema architecture with settable primary and foreign keys. We believe that a fundamental redesign of the core monitoring schema is in order and present our first attempt at that in this paper. Our ultimate goal is to develop a schema that efficiently and accurately captures as much of the monitoring domain as possible and that will extend well with data sets and techniques yet to be developed. For our design, we used the ORM component of Visio Enterprise Architect Edition, and started from the basic object fact relationships (e.g., "an origin has zero or more arrivals"). By using the Visio tool, we are assured that our resultant schema is highly normalized, that all tables have settable primary, foreign and unique keys, and that important constraints are enforced within the database via stored procedures.

RESEARCH ACCOMPLISHED

The core tables are often divided into "Primary" and "Lookup," and we will do the same here. Lookup tables are those that contain static information (i.e., information related to networks, stations, etc.). Primary tables are those that contain dynamic information associated with a given candidate explosion event: location, time, magnitude, which stations recorded it, various amplitude measurements, etc.

To help clarify the discussion below, we introduce the following conventions: table names are in **boldface**, and **c**olumn names are *italicized*.

Primary Tables

Figure 1 shows the Entity Relationship Diagram for the primary tables, with only the key columns (primary = PK, foreign = FK, and unique = UK) shown to simplify. To make our schema more easily understood by those familiar with the current schema, where tables are analogous to tables in the existing schema we have used the same name. Note, however, that none of the tables are exactly the same and some differ significantly, as we discuss below.

Events

An event can be thought of as the highest level object captured by the primary tables (i.e., every other object can ultimately be tied to an event). In the current schema, the **Event** table groups together one or more origins (hypothetical locations), one of which is the preferred origin. However, no origin can be part of more than one event; that is, in the current schema, event-to-origin is one-to-many. This is problematic for a data archive like the NNSA Knowledge Base where events (origin groupings) may have been put together by more than one author. The current design is limiting in another way as well. When multiple origins are relocated simultaneously using location algorithms such as Joint Hypocenter Determination, origin groupings need to be defined. For these reasons, we choose to instead model the event-to-origin relationship as many-to-many, which introduces an additional association table (**EvorAssoc**) that has foreign keys pointing to each of the tables it links. The many-to-many relationship also implies that the **Event** primary key *evid* is no longer needed as a column in the new **Origin** table. With our new association table, an origin can be linked to an arbitrary number of events.

Origins

We have made numerous changes to the **Origin** table for our new schema. Related to the location information represented by an origin, we introduce 3 new tables designed to store specialized information. **AzGap** stores information about station azimuthal coverage, **ModSource** stores information about the source of models used to predict observations, and **RefLocation** stores information about reference locations that an origin may be associated

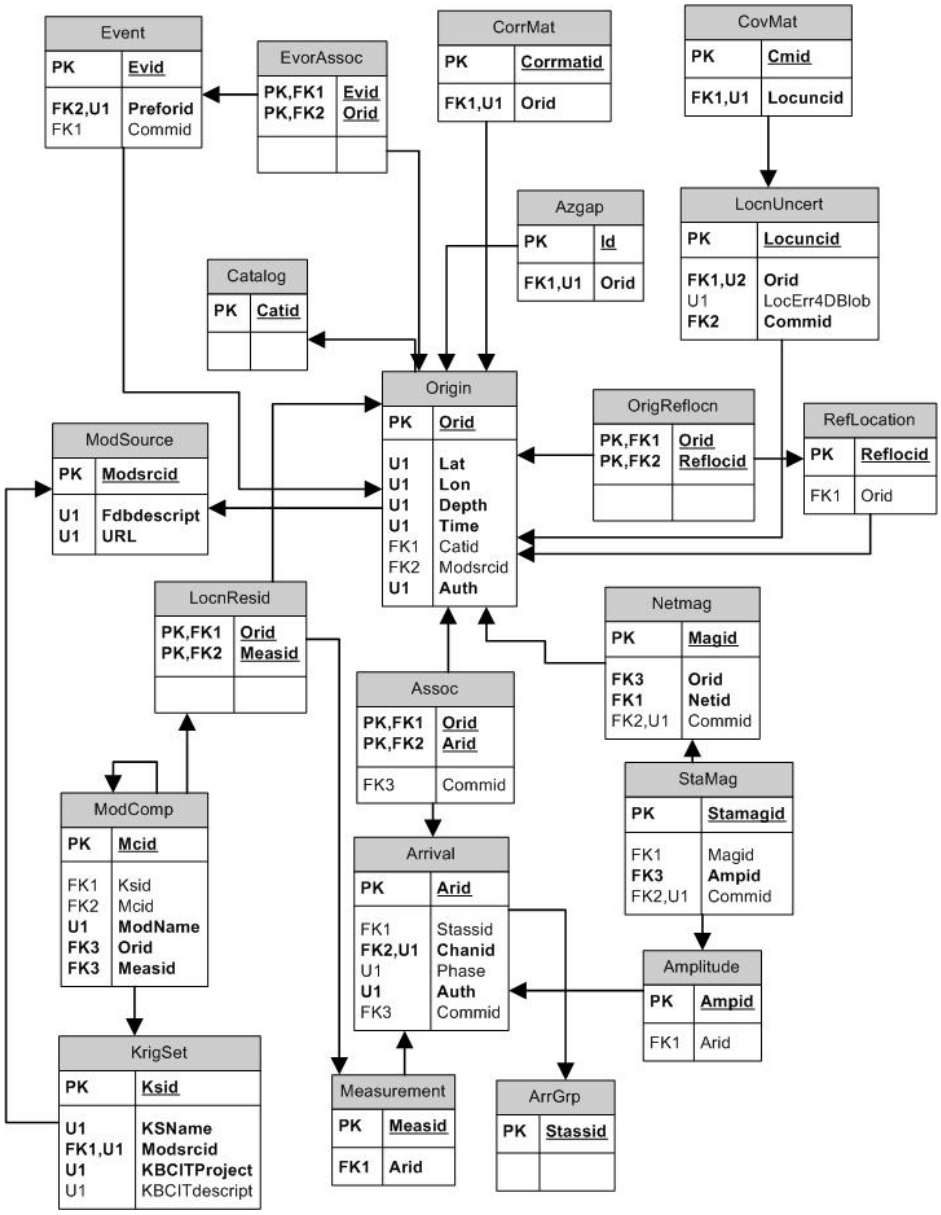


Figure 1. Entity Relationship Diagram (ERD) for the primary tables.

with (e.g., test sites). The **AzGap** ties one-to-one with a specific origin, so this table has *orid* as a foreign key pointing to the **Origin** table. The same **ModSource** information may be used for more than one origin, so this is a one-to-many relationship and hence *modsrcid* (the primary key for **ModSource**) is a foreign key column in the **Origin** table. An origin can have more than one reference location and vice versa, so this is a many-to-many relationship and an association table (**OrigReflocn**) is introduced in the new schema to link **Origin** and **RefLocn**.

Frequently, origins come from specific catalogs, so we have introduced a new **Catalog** table to capture information about the catalogs, including *name*, *description*, and *rank* (*rank* is used to determine which origin in an event is the preferred one). Because **Catalog** to **Origin** is a one-to-many relationship, *catid* (the **Catalog** primary key) is added as a foreign key in the new **Origin** table.

For magnitude information, the changes are all deletions. The current **Origin** table has *mb*, *mbid*, *ms*, *msid*, *ml*, and *mlid*. These ids are all foreign keys to *magid* in the current **Netmag** table, but this is unnecessary because the

Netmag to **Origin** relationship in the current schema is many-to-one, implying that a given origin could have any number of network magnitudes but each network magnitude is related to only one origin (thus *orid* is a foreign key column in **Netmag**). A simple query of **Netmag** by *orid* and *magtype* can fetch mb, ms, or ml values, so these do not need to be stored in **Origin**. Thus, 3 foreign keys in the current **Origin** that point to specific types of magnitudes (*mbid*, *msid*, *mlid*) in **Netmag** are redundant, as are the values associated with them (*mb*, *ms*, *ml*), so we remove all of these in the new schema. If there are multiple network magnitudes of the same type for the same origin, then it might be desirable to designate one of these as the preferred one. To implement this, we have added a new column "*is preferred*" to **Netmag**.

Arrivals

An arrival, also sometimes referred to as a "pick" or "detection," is the fundamental, station/phase-specific observation. Many types of measurements can be made for a given arrival, and one of the most fundamental problems we have found with the current schema is that the information for arrivals is very poorly normalized. An arrival may have travel time, azimuth, and/or slowness observation types measured for it, and the current **Arrival** table has columns to capture information for all three. This is inefficient in that it is very common that not all three types will be measured, so several columns are typically nulled. Further, for each observation type, the basic data to be stored (quantity and uncertainty) are the same, so the columns are highly redundant. Finally, we observe that if another fundamental observation type were deemed to be useful to measure for arrivals, additional columns would have to be added to the **Arrival** table.

We address all of these problems in our new schema by removing the measurement information from **Arrival** and putting it into its own table called **Measurement**. Each measurement has an *obstype*, a *quantity*, an *uncertainty*, and is tied to **Arrival** by the foreign key *arid*. Rows are added to this table only if a particular observation type has been made for an arrival, and there is no limit on how many could be added for a given arrival. In the future, if some new type of observation relevant to an arrival is deemed to be important, the information can be added to the existing tables without having to change the basic structure of any tables.

A further arrival normalization problem occurs with the many amplitude measurements that may be made for a given arrival (peak-to-trough, zero-to-peak, RMS, etc.). The current **Arrival** table has single columns for *amp* and *per*, so these are quickly used. This problem has long been recognized and extension tables have been introduced to capture multiple amplitude measurements per arrival. The USNDC P2B2 schema (Oo et al., 2003) introduces an **Amplitude** table that we have adapted almost exactly for our new schema. This is a generic table with *arid* as a foreign key, so an arbitrary number of amplitude measurements can be made for any arrival. With this table as part of the new schema, we drop the old amplitude support columns in the current **Arrival** table (*amp*, *per*, *SNR*).

Arrivals Linked to Origins

We believe the association of arrival information with origins is also fundamentally flawed in the current schema. Currently, the **Arrival** table and **Origin** tables are linked with a many-to-many relationship via the **Assoc** table. On a high level, that is correct, but if we examine the relationship in more detail, problems emerge.

First, an arrival can be associated with an origin in two ways, and this is not well-modeled. An arrival can simply be tied to an origin in a general sense (it corresponds to a seismic wave caused by the origin), or it can be part of the process that was used to locate the origin. The latter is the reason for the *timdef*, *azdef*, *slodef* columns in the current **Assoc** table. The problem with the current schema is that these two types of information are both put in the same table (**Assoc**). Thus, **Assoc** has columns that are often nulled if an arrival is not used for location at all or is used only for travel time, azimuth, or slowness. Further, as was discussed for the current **Arrival** table, the information recorded in **Assoc** for each type of observation (defining, residual value, and weight) is the same, so the columns are redundant. Finally, though the **Assoc** table might have all three current observation types as defining, and hence using a model to drive location, there is only one column (**vmodel**) to capture the model information, which could easily be different for each observation type. A further problem with the current **Assoc** table is that there is only one *wgt* column when in fact separate weight columns are needed for each observation type. Omission of the weight information for azimuth and slowness observations in the **Assoc** table makes it impossible to recompute the sum squared weighted residuals of the observations associated with an origin if there are any defining azimuth or slowness observations. This is a major flaw since the sum squared weighted residuals are the quantity minimized by the location algorithm that produced the origin to which the assoc entry is linked.

To fix these problems, we made several significant changes. First, we split the information in the current **Assoc** table into two new tables: **Assoc** and **LocnResid**. The new **Assoc** table is basically a subset of the old **Assoc** table, with all the columns relating to locating the origin removed. Retained are columns to hold basic information such as station-to-event distance and azimuth, etc. The removed information is put into the **LocnResid** table, which is a many-to-many link between **Origin** and **Measurement**. This makes sense because the location information is all related to measurements (travel time, azimuth, and slowness) made from the arrival, not the arrival itself. The **LocnResid** table is fully generic, suited to holding information for travel time, azimuth, slowness, or any new observation types that might be introduced in the future.

LocnResid has separate rows for each observation type, so the problem with being limited to a single weight entry and model designation for all observation types is eliminated. Our new schema goes further than this, however, and is able to fully capture the complexity of location models. For many years, the models used for location have actually been composed of a base model plus a set of corrections (elevation, ellipticity, path correction, etc.), so specifying a single model has long proven problematic. Further, in the latest models emerging from location calibration research, the single resultant models developed for a given station/phase are actually composites of several region-specific models blended together. To capture all of this complexity requires a heavily genericized set of tables that can handle arbitrary hierarchies of complexity and we think we have achieved this in our new schema. We capture any of the various model components (base model total or by region, various corrections total or by region) in the new **ModComp** table. **ModComp** has a foreign key pointing to **LocnResid**, so it is easy to find the model components that go with any measurement-origin combination. The hierarchy of model components (e.g., the total base model may come from blending base models from two regions) is captured by endowing the **ModComp** table with a foreign key pointing to itself. To find out if a model component has subcomponents (children), all that is necessary is to query for **ModComp** rows that point to that row. This hierarchy can be extended to as many levels as desired, though we have had no reason to go beyond two (total and regional subcomponents).

To capture the information about the actual source of the models used for a given location (i.e.. file name and location, brief description), we have introduced a new table **ModSource**. Based on our knowledge of the current location models being developed and those anticipated in the future, we expect that all of the model information used to locate an event will come from the same source, so we tie this table to **Origin** rather than **ModComp**. The relationship between **ModSource** and **Origin** is one-to-many, so this introduces the primary key from **ModSource** (*modsrcid*) as a foreign key column in the new **Origin** table.

Currently, only state-of-the-art location algorithms account for correlation between observations, which has been shown to significantly affect both the calculated locations and the uncertainty estimates in some cases (Chang et al., 1983). We have introduced tables to capture this information in the new schema. The individual correlation values are stored in the **CorrEl** (correlation element) table that links together two rows from the **Measurement** table. **CorrEl** rows are in turn grouped into a correlation matrix via the **CorrMat** table. The **CorrMat** table has a one-to-many relationship with **Origin** (i.e., *orid* is a foreign key in **CorrMat**).

Location Uncertainty

As almost no candidate explosion locations are known with perfect accuracy, properly capturing the location uncertainty is essential. The current schema does this with the **Origerr** table, which is a simple extension to the **Origin** table (i.e. it has the same primary key—*orid*—and so just adds additional columns related to location uncertainty). This table captures the full covariance matrix, the one-dimensional (1D) time uncertainty, 1D depth uncertainty and the two-dimensional (2D) horizontal uncertainty ellipse. Three-dimensional (3D) and four-dimensional (4D) uncertainty information is not stored by the current schema. We have introduced separate tables for each type of uncertainty information (**LocErr1D**, **LocErr2D**, **LocErr3D**, **LocErr4D**). The covariance matrix information is stored in its own table (**CovMat**).

Magnitude

We discuss network magnitudes in the Origin section above. Station magnitude measurements are captured in the new **Stamag** table, which is fairly similar to the current **Stamag** table. The changes in this case are deletions. The current **Stamag** table contains numerous redundant convenience columns (*evid*, *orid*, *arid*, *sta*, and *phase*) that can be obtained via joins with other tables, and we have removed these from the new table. If using these joins to access this information proves cumbersome and/or degrades performance significantly, we may reintroduce some of these

fields. We have also given the new **Stamag** its own simple primary key (*stamagid*) rather than the current composite primary (*magid*, *sta*), to make it easier to set up foreign keys pointing from other tables to this one.

Lookup Tables

Figure 2 shows the Entity Relationship Diagram for the lookup tables. Again, to make our schema more easily understood by those familiar with the current schema, where tables are analogous to tables in the existing schema we have used the same name. Also again, however, note that none of the tables are exactly the same and some differ significantly, as we discuss below.

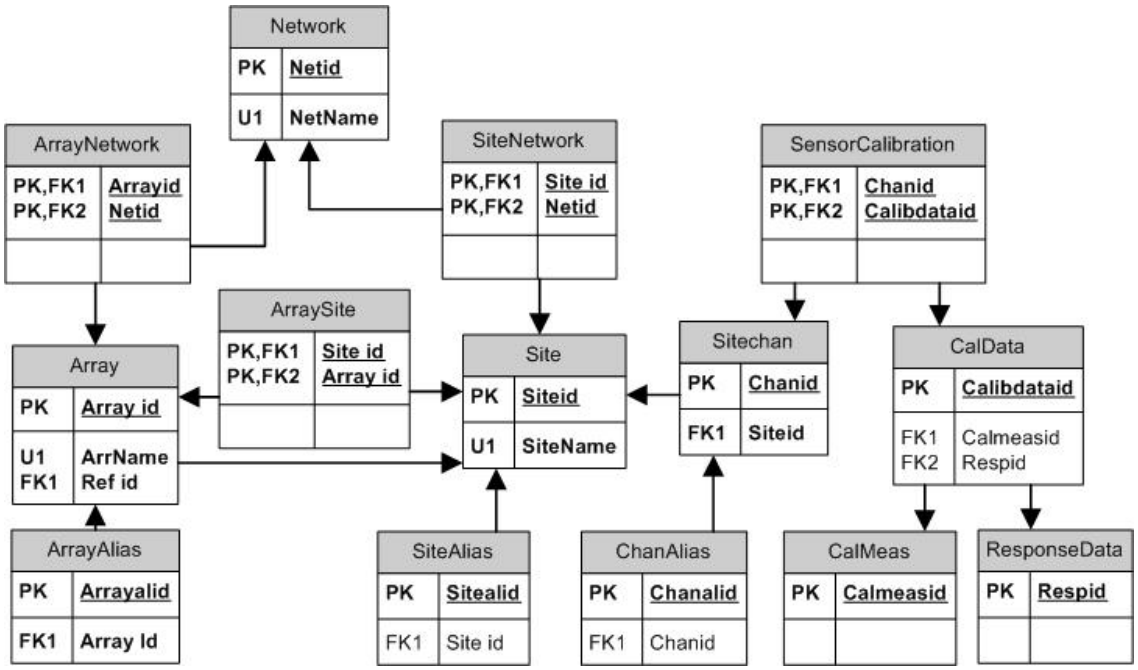


Figure 2. Entity Relationship Diagram (ERD) for the lookup tables.

Network

The top level object for the lookup tables is a network, so we begin there. Our **Network** table is very similar to the current table, except that we add *starttime* and *endtime* to capture the time period during which a network is operational.

Sites

In the current schema, the **Site** table captures the information about the name and location of where seismic instrumentation has been deployed, and we do essentially the same thing. There are however, a few important differences between the new and current **Site** tables. First, we introduce a simple primary key (*siteid*) for the new **Site** table to make it less cumbersome to create foreign keys to point to this table (the current primary key is *sta*, *ondate*). Second, we remove the array-specific columns (*dnorth*, *deast*, *refsta*) from the **Site** table. Storing these in Site is both inefficient because many sites are not part of arrays, and confusing.

Site Groupings into Networks

In the current schema this is done with the **Affiliation** table, and we use a similar table, though we give it a different name—**SiteNetwork**—due to the way we choose to handle arrays. Also, we add *starttime* and *endtime* columns to capture the period during which a site was part of a network. Note that **Network** to **Site** is a many-to-many relationship, so **SiteNetwork** has foreign keys pointing to the primary keys of each of these tables (*netid*, *siteid*).

Arrays

In the current schema, an array is treated just like a network, and so networks of arrays are captured with hierarchies of networks. Though some see this as clever and efficient, we have found it to be confusing and inefficient. In the current schema, for a given array, all elements, including the name given to the array center or beam position (e.g., NORSAR), are rows in the **Site** table. The fact that they are all part of the same array is supposed to be captured in two ways (but seldom is). First, all are given the same *refsta* value in **Site**, which should be the name of the array itself. Second, the array is added as a row to the **Network** table and **Affiliation** rows are added to tie each element to the array. A row is then added to the **Affiliation** table to tie the array to a **Network** (e.g., IRIS). Because the latitude, longitude information stored with each site may not be accurate enough for array processing (e.g., FKs), offsets relative to the array center are stored in the **Site** table (*dnorth*, *deast*). As pointed out above, this is inefficient because all site rows carry these columns, even though most sites are not array elements. A further problem is that if a site is to be used as an element of more than one array (easily done given that arrays are virtual entities) only one set of offsets can be stored.

In our new schema, we take a much more direct approach and model arrays according to the simple fact statements "arrays consist of one or more sites" and "zero or more arrays may be associated with a network." In our approach, there is an **Array** table that captures the basic information about the array (e.g., name, and a foreign key pointing to **Site** to indicate which site position is the center or beam position). **Array** to **Site** is a many-to-many relationship, so a new association table **ArraySite** is introduced. This is the table where the offsets (*dnorth*, *deast*) are stored, as well as the *starttime* and *endtime* of each site's membership in the array (which is not captured at all in the old schema). The many-to-many array membership in networks is captured with another association table, **ArrayNetwork**. Again, this table has *starttime* and *endtime* columns to capture the period of membership for each array in a network.

Channels

Each site is composed of one or more seismic instruments to record the motion of the Earth. Information about each instrument deployed at a site is captured in the current schema with the **Sitechan** table, and our **Sitechan** table is nearly identical. One important difference is that we explicitly model the one-to-many relationship between **Site** and **Sitechan** by giving our **Sitechan** table a foreign key pointing to the primary key in **Site** (*siteid*). We use *starttime, endtime* columns in **Sitechan** to record the operational period for the particular channel, which may be different from the *starttime, endtime* appropriate for the site as a whole.

Aliases

A further source of confusion with lookup information not captured by the current schema is that arrays, sites, and channels can all have aliases (i.e., synonyms). For example, a short period vertical channel may be SHZ or SZ. The fact that such names are equivalent is not captured in the current schema, and this can lead to errors when using data. We fix the problem by introducing a set of very similar tables (**ArrAlias**, **SiteAlias**, **ChanAlias**) to link the various aliases to the "true" name. Each of these tables has *starttime* and *endtime* columns to capture the duration of the alias.

Mapping the Current Schema to the New Schema

Should our proposed schema be chosen for use with real monitoring systems, large amounts of old-format data would need to be converted to the new schema. This is easily done, as we show in Figure 3 for a few of the key tables.

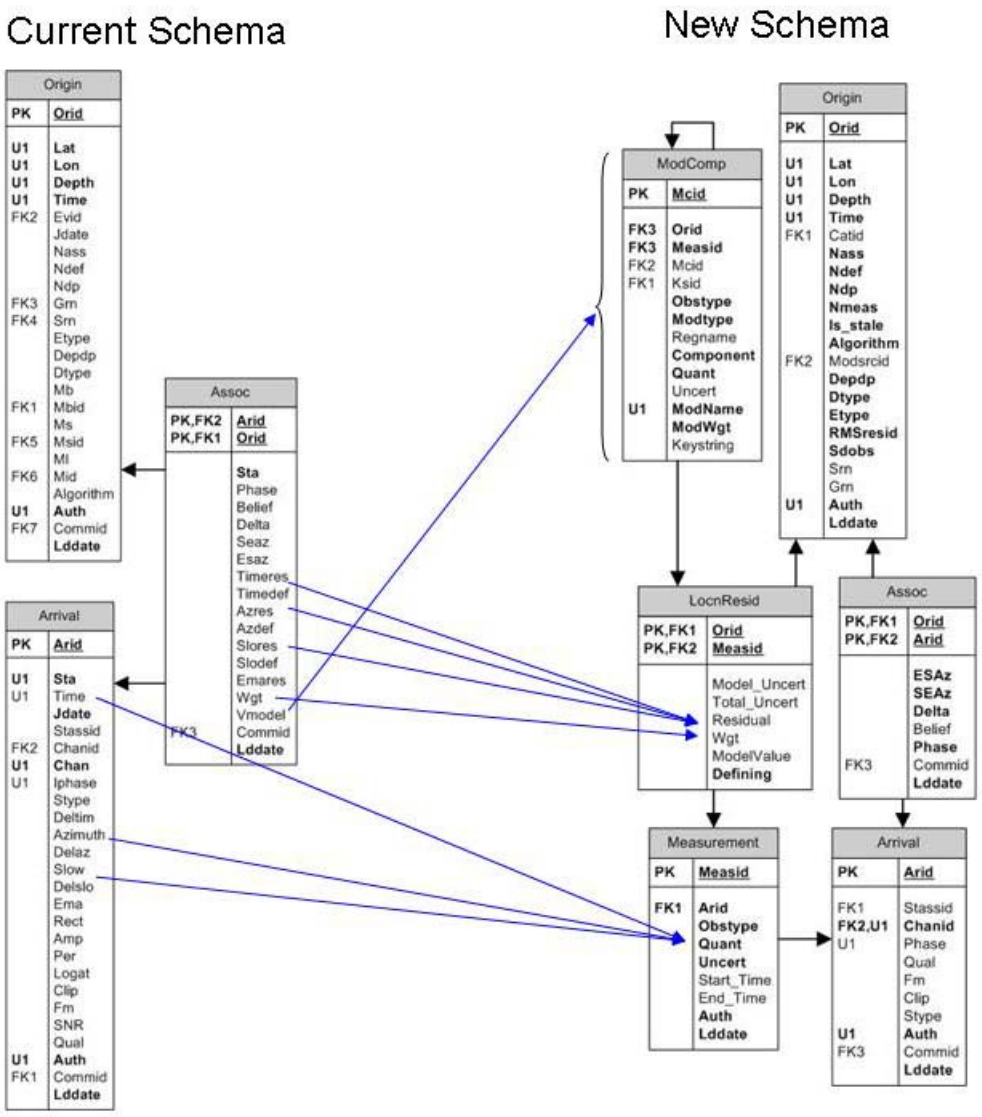


Figure 3. Remapping current schema tables to new schema tables.

In this example, we show how an origin and associated arrivals used for location would be converted from the current schema to the new schema. Much of the information in the current **Arrival** table (*phase*, *qual*, *fm*, etc.) would map directly to the new **Arrival** table. Travel time, azimuth, and slowness observations and estimated uncertainties (*time*, *deltim*, *azimuth*, *delaz*, *slow*, and *delslo*), however, would be mapped to separate rows in the new **Measurement** table. Information from the current **Assoc** table would be mapped to 3 tables in the new schema: **Assoc**, **LocnResid**, and **ModComp**. Residual values for travel time, azimuth, and slowness (*timeres*, *azres*, *slores*) would be mapped to separate rows in the **LocnResid**. The same would be true for the weights for each observation type, except that the current schema only stores one of these (*wgt*), so we would have to specify which observation weight to map to in the new schema. The woefully inadequate *vmodel* column in the current **Assoc** table essentially gets remapped to the entire **ModComp** table, though there is not enough information available for historic events to properly populate the new table. Presumably the string in *vmodel* could be mapped to *modname* in the **ModComp** table. The remaining information in the current **Assoc** table (*esaz*, *seaz*, *phase*, *belief*) maps directly to corresponding columns in the new **Assoc** table.

Similar remaps can be made for the rest of the information in the current schema, and using these remaps we have written some simple scripts to translate information archived in the old format schema into the new schema.

Unfortunately, by using the current schema in the first place, some critical information was never stored properly (e.g., detailed information about the models used for location), so the new tables will not be fully populated for historical data. There is still a gain in doing the translation because the new schema is more efficient and more amenable to code development, but the full benefit of our data model can only be realized when it is used from the start to archive data.

CONCLUSIONS AND RECOMMENDATIONS

We have designed a new database schema to manage information related to nuclear explosion monitoring that captures all of the information stored with the current schema but overcomes many of the problems with the current design. Guiding principles to which we sought to adhere in our design include

- all tables that store fundamental quantities (**Origin**, **Arrival**, **Site**, etc.) should have single, numeric, primary keys;
- having columns that are very frequently null in a table is to be avoided whenever possible; and
- the schema should form a cohesive, integrated entity with settable primary, unique and foreign keys.

Our design

- allows **Origins** to be grouped together in multiple different ways to reflect **Events** defined by different researchers/institutions and to allow the definition of groups of **Origins** that were located simultaneously using Joint Hypocenter Determination;
- properly normalizes magnitude information;
- adds tables that document the source of earth model information and the various components that constitute the predictions of observed quantities;
- defines the **Measurements** made on an **Arrival** separately from the **Arrival** itself, thereby reducing redundancy and enhancing extensibility;
- associates **Measurements** with **Origins** using **Residuals** in addition to the currently defined association between **Origin** and **Arrival**; and
- defines **Sites**, **Arrays** and **Networks** in a way that is more intuitive and that promotes discipline.

Our schema is backward-compatible with the current schema in that it is possible to capture all information stored in the current schema, and views of the new schema can be constructed which allow current software to access information stored in the new schema.

ACKNOWLEDGEMENTS

We thank Richard Stead, Julio Aguilar-Chang, Stan Ruppert, Dorthe Carr, and Mark Harris for countless discussions about problems associated with the current schema and how they should be fixed.

REFERENCES

Anderson, J., W. Farrell, K. Garcia, J. Given, and H. Swanger (1990), Center for Seismic Studies Database: Version 3.0 Structure, SAIC document.

Carr, D. (2004), National Nuclear Security Administration Knowledge Base Core Table Schema Document, Sandia National Laboratories, SAND2002-3005.

Chang, A. C., R. H. Shumway, R. R. Blandford, and B. W. Barker (1983), Two Methods to Improve Location Estimates—Preliminary Results, *Bull. Seismol. Soc. Am.* 73: 281–295.

Oo, K., D. Irons, A. Henson, and C. Morency (2003), Database Design Description (DBDD) Phase 2, Rev C, United States National Data Center (US NDC), SAIC-01/3047 (REV C).